2018-07-30

RIPPLiT and ChimeraTie: High throughput tools for understanding higher order RNP structures

Mihir Metkar
University of Massachusetts Medical School

RIPPLIT AND CHIMERATIE: HIGH THROUGHPUT TOOLS FOR UNDERSTANDING HIGHER ORDER RNP STRUCTURES

A Dissertation Presented By

MIHIR METKAR

Submitted to the Faculty of the University of Massachusetts Graduate School of Biomedical Sciences, Worcester in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY

July 30, 2018

Molecular Biology

RIPPLIT AND CHIMERATIE: HIGH THROUGHPUT TOOLS FOR UNDERSTANDING HIGHER ORDER RNP STRUCTURES

A Dissertation Presented By

MIHIR METKAR

This work was undertaken in the Graduate School of Biomedical Sciences
Integrated Graduate Program

Under the mentorship of

Melissa J. Moore, Ph.D., Thesis Advisor

Job Dekker, Ph.D., Thesis Advisor

Athma A. Pai, Ph.D., Member of Committee

Thoru Pederson, Ph.D., Member of Committee

Scot A. Wolfe, Ph. D., Member of Committee

Karla Neugebauer, Ph. D., External Member of Committee

Victor R. Ambros, Ph. D., Chair of Committee

Mary Ellen Lane, Ph.D., Dean of the Graduate School of Biomedical Sciences

July, 30 and 2018

DEDICATION

This thesis is dedicated to my parents, Ashok and Anjushree for their support and encouragement.

My grandparents Vasant and Manorama, for their love and being my first teachers.

My lovely but troublesome wife, Vahbiz without whom I could never have achieved this.

ACKNOWLEDGEMENTS

The last few years have been an incredible learning experience and with heartfelt gratitude I would like to acknowledge all the people without whom this work would not have been possible. First of all, I wish to express my sincere gratitude to my thesis advisors and mentors, Melissa Moore and Job Dekker for their continuous guidance, unwavering support and for being amazing human beings. I first joined Melissa's lab and her mentorship was instrumental in shaping my scientific temper. Subsequently, I became a joint student in Job's lab which further sharpened my scientific acumen. They gave me the freedom to explore new horizons and at the same time always made themselves available for insightful discussions and kept me focused. Over the years, they improved my writing, presentation skills and helped me become a better scientist as well as a better person.

I have been very fortunate to have the guidance of a supportive committee: Oliver Rando, Victor Ambros, Scot Wolfe, Andrei Korostelev, Thoru Pederson, and Athma Pai. Apart from their scientific brilliance, insights and valuable advice, they have been very approachable, pragmatic and have always been there to encourage and motivate me. I would also like to thank Karla Neugebauer for agreeing to be the external examiner for my defense and for a stimulating discussion. A special thank you to Kundan Sengupta for encouraging me to pursue research and a PhD.

I owe a special thanks to Guramrit Singh and Emiliano Ricci who were my postdoctoral mentors in the Moore lab, and Bryan Lajoie and Hakan Ozadam who were my bioinformatics mentors from the Dekker lab. Their support and training over the years, has been instrumental in shaping my thesis project. I would also like to specially thank Rina Paladino, Melissa's administrative assistant, without her organizational skills, and timely reminders, PhD life would have been very difficult. I would like to thank Moore lab members- Weijun Chen, Kelly Lemoncelli, Blandine Mercier, Akiko Noma, Rachel Niederer, Jing Yan, Christian Roy, Erin Heyer, Eric Anderson, Lingtao Peng, Alicia Bicknell, Laureen Murtha, Abby Smith, Alper Kucukural, Joerg Braun, Andrew Franck, Carrie Kovalak, Harleen Saini and Ami Ashar; Dekker lab members- Nicki Fox, Liyan Yang, Ye Zhan, Johan Gibcus, Ankita Nand, Sergey Venev, Erica Hildebrand, Yu Liu, Anne-Laure Valton, Allana Schooley, Houda Belaghzal, Marlies Oomen, Kristen Abramo, Betul Akgol Oksuz and Bastiaan Dekker for creating a wonderful, fun and friendly learning environment, and for stimulating discussions and honest, constructive feedback.

I would like to thank my wonderful friends in the US and back at home for their encouragement and support. A special thank you to Satyajeet, Pratik, Gautam, Anuraag and Akshay for always being there for me despite the distance. I want to take this opportunity to thank Ami-Hemal, Ankit-Pallavi-Soham-Tanishi, Sneha-Varun, Sandhya-Karthik-Haasini, Swapnil-Nishgandha, Satya-Harshada Pallo, Sreya, Ankita, Mayuri, Ashwin-Neha and all my cricket friends (Worcester Speedsters) for being like family away from home. Life in UMass has been an

adventurous ride thanks to all the vacations, camping trips, hikes, birthdays and festivals celebrated together. I especially am grateful for the wonderful memories, love and support through the years.

A heartfelt gratitude to my parents, Ashok and Anjushree for their love, continuous support, encouragement and for never losing faith in me. They have taught me the importance of education and the value of learning new skills. I am also indebted to my grandparents, Vasant and Manorama, for their love and taking care of me through the initial years of my life. They taught me to give my best in any task I undertake.

Finally, I owe my deepest gratitude to my wife, Vahbiz, for being my closest friend and biggest critic. I cannot thank her enough for dealing with my quirks, tolerating my mischiefs, for always being there for me and for helping me improve myself. I feel extremely lucky to have her in my life.

Words aren't enough to express my gratitude towards everyone who has helped me become who I am. I am extremely lucky and blessed to have them all in my life and wouldn't be able to achieve any of this without their belief, love and support.

ABSTRACT

Even after their discovery more than 60 years ago, little is known about how messenger RNAs (mRNAs) are packaged inside the cells. To ensure efficient and accurate delivery of the intended message to its proper destination, it is important to package the informational molecule in a way that protects it from premature degradation but also proper decoding at the destination. However, very little is known about the this fundamentally important step of mRNA packaging inside eukaryotic cells. To this end, we developed a novel approach, RIPPLiT (RNA ImmunoPrecipitation and Proximity Ligation in Tandem), to capture the 3D architecture of the ribonucleoprotein particles (RNPs) of interest transcriptome-wide. To begin with, we applied RIPPLiT to the exon-junction complex (EJC), a set of proteins stably bound to a spliced RNA. EJCs have been shown to interact with other proteins like SR- and SR-like to form megadalton sized complexes and help protect large regions of mRNAs. Thus, we hypothesized that these RNPs would provide an ideal system to elucidate the higher order organization of mRNPs.

Preliminary analysis of data obtained from RIPPLiT consisted of “chimeric reads”, reads with multiple RNA fragments ligated together, which could not be analyzed with any of the existing bioinformatics tools. Thus, we developed a new bioinformatics suite, ChimeraTie, to map, analyze and visualize chimeric reads. Performing polymer analysis on chimeric reads obtained for hundreds of mRNAs,

we were able to predict that mRNPs are linearly and densely packed into flexible rod-like structures before they undergo translation.

In this thesis, along with the detailed biological conclusion, I have also provided a step-wise manual to perform RIPPLiT experiment and analyze the ensuing data using ChimeraTie.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION AND NOMENCLATURE

RNP	Ribonucleoprotein
hnRNP	Heterogenous nuclear RNPs
mRNA	messenger RNA
tRNA	transfer RNA
rRNA	ribosomal RNA
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
ncRNA	non-coding RNA
miRNA	micro RNA
dsRNA	double-strand RNA
ssRNA	single-strand RNA
RBP	RNA-binding protein
EJC	Exon junction complex
XIST	X-inactive specific transcript
AU	Adenylate-Uridylate
ARE	AU-rich elements
TTP	tristetraprolin
HuR	Hu Antigen R
YBX1	Y-Box binding protein 1
PABP2	Poly-A binding protein 2
eIF4AIII	Eukaryotic initiation factor 4A-III
PNK	Polynucleotide kinase
T4 Rnl	T4 RNA ligase
CIP	Calf intestinal phosphatase
AMT	4'-aminomethyltrioxsalen
ATP	Adenosine triphosphate
DTT	1,4-Dithiothreitol
IP	Immunoprecipitation
EM	Electron microscopy
PAGE	Polyacrylamide gel electrophoresis
DMS-seq	Dimethyl sulphate treatment to a massively parallel sequencing readout
SHAPE	Selective 2'-hydroxyl acylation analyzed by primer extension
CRAC	Cross-linking and analysis of cDNAs
CLASH	Cross-linking and sequencing of hybrids
hiCLIP	RNA hybrid and individual-nucleotide resolution ultraviolet

	cross-linking and immunoprecipitation
LIGR-seq	Ligation of interacting RNA followed by high-throughput sequencing
SPLASH	Sequencing of psoralen cross-linked, ligated, and selected Hybrids
PARIS	Psoralen analysis of RNA interactions and structures
MARIO	Mapping RNA-RNA interactome
RPL	RNA Proximity Ligation
RIPiT	RNA immunoprecipitation in tandem
RIPPLiT	RNA immunoprecipitation and proximity ligation in tandem
HLB	Hypotonic lysis buffer
DLB	Denaturing lysis buffer
IsoWB	Isotonic wash buffer
DWB	Denaturing wash buffer
PIC	Protease inhibitor cocktail
5′	5 prime end of the polynucleotide
3′	3 prime end of the polynucleotide

COPYRIGHT INFORMATION

Figure 2.9D has been adapted from the following publication

Lu, Z., Zhang, Q.C., Lee, B., Flynn, R.A., Smith, M.A., Robinson, J.T., Davidovich, C., Gooding, A.R., Goodrich, K.J., Mattick, J.S.*, et al.* (2016). RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. Cell *165*, 1267-1279.

License Number	4303750134010
License date	Mar 07, 2018
Licensed Content Publisher	Elsevier

Chapter I

Introduction

The central dogma of cell biology states that the genetic information encoded in our “DNA” is decoded to its functional form “protein” (Crick, 1958). This process entails first transcribing the genetic information to a messenger molecule- “messenger RNA” (mRNA), from where it carries this information to a target destination where it is decoded to proteins in a temporal and spatial manner. mRNA molecules that code for proteins are key components that facilitate the transmission of genetic information in all life forms- from viruses, to higher order complex mammalian organisms. However, the challenge lies in the faithful transport and reproducibility of the information encoded in mRNAs at the desired destination. To this end, mRNAs need to package in a way that not only protects its integrity (protection from premature degradation) but also ensures its proper decryption (targeted translation).

Need for RNP packaging

Principles underlying information transfer are universally applicable irrespective of the nature of information (Gatenby and Frieden, 2007). For instance, coding theory in mathematics that deals with efficient and accurate transfer of information (telephone call) from a source to a desired destination describes how “noise” (e.g., competing signal, poor speech, poor hearing, random disturbance) can affect the “transferred message” (Ash, 1990). Inside our cells, informational transfer, which in this case happens through mRNAs demonstrate a

remarkable ability to bypass the noise (Cellular noise can be defined as i. steric hindrance of nuclear pores for exit of mRNA from the nucleus ii. protection from degradation by endogenous nucleases or iii. protection from premature translational signals) (**Figure 1.1**). This, in part, can be attributed to how the RNA is packaged inside the cell, and thus the aim of this thesis is to understand some of the fundamental principles of mRNA packaging.

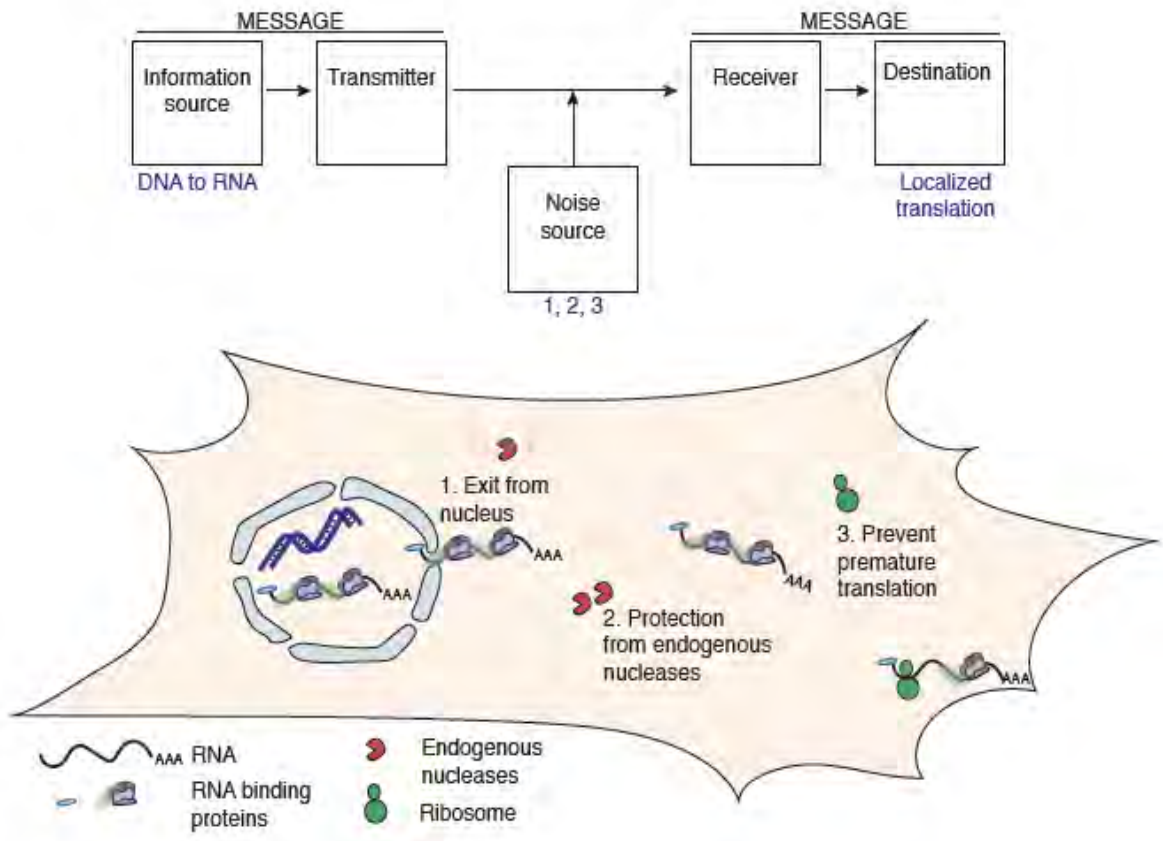


Figure 1.1. Universal principles of Information transfer

An illustration of the stages of information transfer in coding theory and its correlation to information transfer inside the cells. Information from the source (DNA) will have to bypass various levels of noise (protect the mRNA) before it reaches its destination (for localized translation). Model of a communication system (top) is adapted from (Ash, 1990).

Universal properties of packaged and transmitted information

Consistent with all information transfer systems, the message is only one third of the final product. An information transfer theory (Landau et al., 1982) defines “Product” of information transfer as,

Product = Subject content + User + Package

Thus, the final “Product” of the information in our genetic code is a combination of “Subject content” (mRNA), “User” (translational machinery, location of translation) and “Package” (mRNA packing by the associated proteins). Packaging, according to this information transfer model, is the most complex and least understood of the factors (Landau et al., 1982). Not surprisingly, similar constraints lie in messenger RNA transfer and packaging as well.

Components of information package

An information “package” consists of 2 parts- first is the actual piece of information that is being preserved or transported (e.g., image file on hard disk, here the mRNA molecule being packaged) and the second is the associated representation information (e.g., meta-file for the image, here the RNA-binding proteins associated with the mRNA) that are necessary to make the information understandable to the designated user (e.g., DVD player, nuclear pore complex, translational machinery).

The RNA associated proteins play multiple different roles during the various steps of the life-cycle of an RNA (Moore, 2005; Singh et al., 2015) (**Figure 1.2**). Some of which are as follows-

1. They describe the information source including the processing history. In the case of an mRNA, one could imagine the exon-junction complex which is deposited after splicing as a marker (Le Hir et al., 2000; Woodward et al., 2017).
2. Some proteins can act as protective shield that protects the information content from unnecessary alteration. Purified YB-1 was shown to coat RNAs in a non-sequence specific manner and package them *in vitro* (Skabkin et al., 2004).
3. They provide a set of identifiers to uniquely identify the particular information content. This could be a set of sequence specific proteins that bind to particular mRNAs. Exemplifying this is Adenylate-Uridylate-rich elements (AU-rich elements; AREs)), which are AU rich sequence motifs in RNAs that bind to specific antagonistic proteins like hnRNP D, and Hu Antigen R (HuR) and regulate mRNA degradation. Binding of hnRNP D to AU-rich elements leads to recruitment of exosome (RNA degradation machinery) and degradation while binding to HuR to ARE elements antagonizes degradation (Brennan and Steitz, 2001).
4. They describe how the information relates to other information outside the package (See example in point 3).

5. Some proteins provide the information for access rights (preservation, distribution or usage) for the information content. Stau1 was shown to be required for proper localization of bicoid mRNAs to the anterior pole of *Drosophila* egg (Ferrandon et al., 1994).

Taken together, both (RNA + proteins) these information domains make one complete information molecule. Thus, to correctly understand the steps in the life-cycle of an mRNA, it is important to study it in context of the overall information content as well as the description information i.e., mRNA in association with its associated proteins.

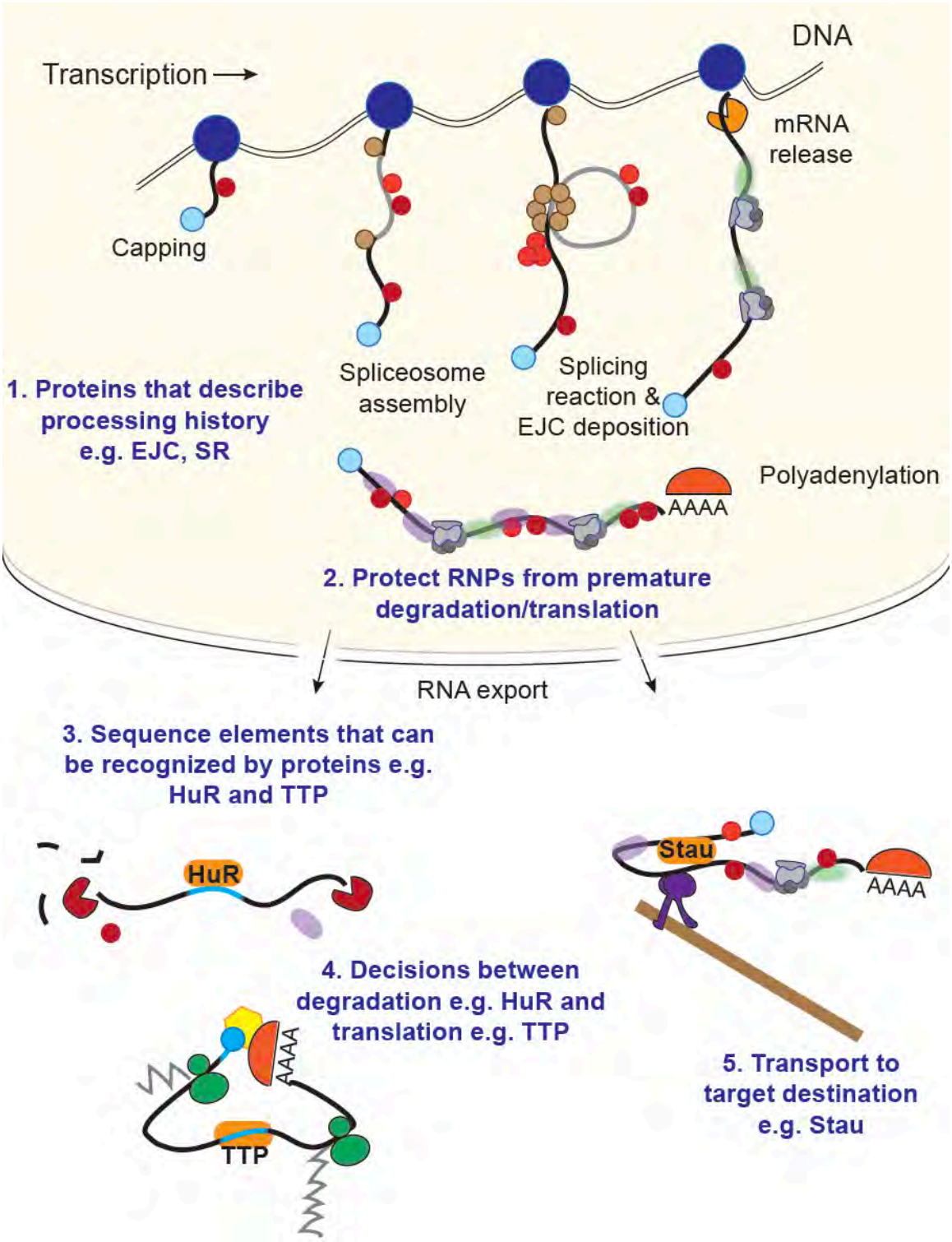

DNA
Transcription
mRNA release
Capping
Spliceosome assembly
Splicing reaction & EJC deposition
1. Proteins that describe processing history e.g. EJC, SR
Polyadenylation
AAAA
2. Protect RNPs from premature degradation/translation
RNA export
3. Sequence elements that can be recognized by proteins e.g. HuR and TTP
HuR
Stau
AAAA
4. Decisions between degradation e.g. HuR and translation e.g. TTP
AAAA
TTP
5. Transport to target destination e.g. Stau

Figure 1.2. mRNA and RNA-binding Proteins together constitute a packaging unit

An illustration of the life cycle of an mRNA with different proteins associated with it. Proteins associate with mRNAs from birth to death, regulating them through their life cycle.

Role of RNP structure in the life cycle of RNAs

Similar to DNA and proteins, RNAs adopt certain 3D organization inside cells. Given that RNAs are single-stranded, they can form multiple types of 2D structures like hairpins, bulges, loops and pseudoknots, which in turn lead to complex and diverse 3D structures. 3D conformations of highly abundant ncRNAs like rRNAs, tRNAs and snoRNAs (Palazzo and Lee, 2015) have been extensively studied. Since, most of these RNAs perform structural or enzymatic functions, they are required to have conserved 3D structures (Rivas et al., 2001). Thus, due to their abundance and near similar higher order structures, obtaining high resolution 3D structures was possible. However, compared to these, mRNAs do not form stable 3D structures. Further, there are ~10^5 mRNA molecules within a cell (Palazzo and Lee, 2015) made up of 1000's of different mRNA species. Thus, the variability in mRNA structures and their diversity has made it difficult to study mRNA higher order structure inside cells. Although very few studies have looked at 3D organization of mRNAs, there are plenty of studies that demonstrate the importance of mRNA secondary structure throughout the life cycle of an mRNA.

1. Co-transcriptional folding affects transcriptional and translational decision

   From birth of the RNA, it undergoes conformational changes (or folding) which mediate various stages of its life cycle. As soon as the nascent transcript is synthesized, RNAs begin to fold co-transcriptionally in a directional manner

from 5' to 3' end (Lai et al., 2013). This process of co-transcriptional folding has been well studied in the case of riboswitches, which are regulatory elements often located in the 5' UTR of bacterial mRNA. Riboswitches can respond to metabolites or metal ions in the environment, alter their conformation and thereby signal for transcription attenuation or translation initiation of the same mRNA they are coded in (Lai et al., 2013). A recent study performed SHAPE-seq to study co-transcriptional folding of a fluoride riboswitch, and through their study they identified the precise structure at which the nascent RNA mediates a genetic decision (Watters et al., 2016).

2. Secondary structure affects alternative splicing of genes

Although there are ~38,000 genes in the human genome, they code for ~152,000 transcripts (NCBI Homo sapiens Annotation Release 108). During the process of splicing of a pre-mRNA mediated by the spliceosome, introns are removed while exons are joined together. However, by changing what parts of the pre-mRNA are removed or included (alternative splicing), a great diversity in mRNA sequence can be achieved. Approximately 95% of the human genes are alternatively spliced (McManus and Graveley, 2011) giving rise to multiple isoforms for the same gene. Multiple factors affect this process like proteins that act as splicing enhancers or repressors, strength of splice sites and also secondary structure of the RNA. For instance, local secondary structure near the 5' or 3' splice site can lead to the formation of base-pairing

interactions between the splice sites and the proximal sequences. This can prevent the spliceosome from recognizing these sites (McManus and Graveley, 2011) and therefore alternatively splice the RNA. A classic example of sequence, structure and function relationship is the microtubule-associated protein tau (MAPT or tau) pre-mRNA. Specific mutations in the tau gene destabilizes a hairpin structure at exon 10-intron10 junction (Liu and Gong, 2008). This altered structure interferes with the interaction of the transcript with U1 snRNP, thereby resulting in alternative splicing.

Secondary structure can affect splicing not only through short range base-pairing interactions, but also longer-range contacts. This is best exemplified by the *Drosophila* Dscam gene (Celotto and Graveley, 2001; McManus and Graveley, 2011). This particular gene can give rise to >38,000 isoforms through alternative splicing. The introns of this gene consist of conserved sequences that can forms base-pairs with regions thousands of nucleotides away. By modulating these base-pairs, the gene can include or exclude certain exons and thus give rise to multiple isoforms.

3. Exit from the nucleus

Few studies have looked at how RNP structure mediates the nuclear export process. *In vivo* live imaging study of single labelled mRNPs during nucleocytoplasmic transport demonstrated that the length of an mRNA does

not affect the rate of transport of the mRNP. This led to the speculation that mRNPs maybe compacted in a similar manner before their export (Katahira, 2015; Mor et al., 2010).

EM Structural studies from Balbiani ring mRNPs in the salivary glands of *Chironomus tentans*, demonstrated that the ultra large mRNP undergoes a conformational change from a globular structure to an rod-shaped one as they are threaded out of the nuclear pore complex (Skoglund et al., 1983).

4. RNA editing by ADAR requires specific secondary structure in its targets

A-to-I editing is the most common form of RNA base editing found in humans (Bajad et al., 2017). RNA editing may affect the base-pairing strength of the edited nucleotide or affect the binding an RBP to this position thus leading to structural and functional changes. Additionally, during translation, I is read as guanosine and thus A to I can alter protein sequence and therefore function. Furthermore, the A-to-I editing target sites are present in clusters, thus could lead to large structural changes in the target RNA (Solomon et al., 2017).

A-to-I editing is performed by a family of proteins called ADAR (adenosine deaminases acting on RNA). ADAR proteins have been shown to require specific RNA secondary structural features to perform base-editing (Tian et al., 2011). The ADARs recognize double-stranded RNA regions formed by inverted

repeat sequences in RNAs as substrate for A-to-I editing (Solomon et al., 2017).

5. Secondary structural elements can play an important role in mRNA localization

mRNA localization to specific compartment of a cell is a significant mode of regulation gene expression. Spatially segregating mRNAs can ensure that the protein can only be made at the target destination. One of the best studied examples of this process is the Bicoid mRNA localization to the anterior pole of *Drosophila* oocyte (Ferrandon et al., 1994). By localizing specific mRNAs to specific locations, the oocyte creates protein gradients that help in regulating zygote specific genes at those target regions. However, proper localization of the bicoid mRNA to the posterior end requires a stem-loop structure in the 3' UTR of the mRNA that is bound by Staufen protein. Furthermore, these loops from different bicoid mRNAs multimerize to form larger particles that are then transported to the anterior end (Ferrandon et al., 1997). This example demonstrates the importance of both intramolecular as well as intermolecular base-pairing structures in the proper functioning of an mRNA's life-cycle.

6. RNA translation is affected by the location of the secondary structure within it

Secondary structural elements affect the process of translation in multiple different way. A secondary structure element (e.g., a stable hairpin), near the start codon creates physical hindrance for the ribosome and thus reduces translation efficiency. Similarly, increased structure in the 5' UTR adversely affects the mRNA's translation (Ding et al., 2014; Kramer and Gregory, 2018). On the other hand, structure in the coding region correlated positively with ribosome association (Li et al., 2012). In addition, RNA structure probing methods also observed higher structure *in vivo* in the region 50 nt upstream of alternative polyadenylation sites (Ding et al., 2014). Thus, secondary structural elements can affect translation in different ways depending upon their location and strength within a transcript.

7. Secondary structure regulates RNP granule formation

The crowded environment of cell provides a challenge to spatially and temporally separate different processes. One way to achieve this inside the cells is by forming membrane-less organelles with high local density of factors (e.g., RNA and proteins) that can be cycled rapidly (Banani et al., 2017). These organelles are thought to form liquid-liquid phase separations which allows them to maintain distinct organization without a solid separation. An example of such an organelle is Stress granules formed in the cytoplasm upon inhibiting

translation (Van Treeck et al., 2018). It was proposed that upon translational block, translationally inhibited mRNP can self-assemble into Stress granules and thus regulate translational response to stress. It has been demonstrated that RNA-RNA interactions play a pivotal role in assembling the Stress granules.

Recently, another study looked at two types of granules formed in the filamentous fungus *Ashbya gossypii* which asssemble based on the mRNA present in these granules (Langdon et al., 2018). Further, they demonstrated that secondary structure of the component mRNA regulates which granules this particular mRNA will incorporate with. Thus, this study clearly showed that RNA secondary structure can play an important role in regulating assortment of an RNA into liquid-liquid phase separations.

All these examples demonstrate the importance of RNA structure (here, secondary structure) in regulating various cellular processes. However, compared to mRNP 2D structure, regulation of mRNP through inherent 3D packaging has largely been overlooked. Given the importance of secondary structure in mRNP life-cycle, mRNP 3D structure will undoubtedly be equally significant. Thus, the first step in deciphering the effect of mRNP 3D structure is to under the 3D organization of the messages.

Early studies on mRNP packaging

Earliest attempts to understand the RNP packaging came from electron microscopy (EM) studies done on pre-mRNAs (Dreyfuss, 1986). Heterogenous nuclear RNPs (hnRNPs) were isolated from nuclei, treated with mild RNases followed by separation using density centrifugation. The undigested nuclear hnRNP fraction sedimented from 30 to ~200S while the RNase treated hnRNPs sedimented at 30-40S (Dreyfuss, 1986; Samarina et al., 1968). It was proposed that hnRNAs have a nucleosome like arrangement inside the nuclei with each monomer comprising ~700 nt of RNA wrapped around a core of hnRNP (A1,A2,B1,B2,C1,C2) proteins (Conway et al., 1988) akin to DNA wrapped around histones with each monomer connected by a small piece of RNA bridging the adjacent monomers. These structures were termed as ribo-nucleosomes and the model was called “beads-on-a-string”. The presence of ribo-nucleosomes was validated using imaging of hnRNPs under EM studies that showed hnRNAs packaged as clusters of small spherical structures of relatively equal size, presumably, the 30S monomers clustering to form the whole hnRNP (Malcolm and Sommerville, 1977; Samarina et al., 1968). These ribo-nucleosomes could be visualized co-transcriptionally and dispersed into individual spherical (30S) particles upon treatment with RNases (Dreyfuss, 1986; Samarina et al., 1968). Further, hybridizing 30S purified hnRNA with cytoplasmic RNA yielded only 5-10% hybridization (Kinniburgh and Martin, 1976). However, almost all of the cytoplasmic RNA was represented in the 30S nuclear hnRNA, indicating that the hnRNPs

consist of pre-mRNAs. Given that mRNAs are 16-fold shorter than pre-mRNAs (human genome annotation, average length of gene = 48, 598 nt and average length of mRNAs = 3,437 nt), it was difficult to visualize spliced packaged mRNPs.

Another study assembled a specific RNA (of differing lengths) with a purified YB-1 protein, which is known to bind an RNA throughout its length, *in vitro*. Interestingly, even this complex organized as “beads-on-a-string” when visualized under EM and was shown to compact the RNA 5-fold inside the RNP (Skabkin et al., 2004).

However, very few studies have looked at the *in vivo* packaging of spliced mRNAs. One such study done on relatively short (~1,250 nt) yeast mRNAs, enriched mRNPs by pulling down NAB2, a polyA binding proteins and then size-separated the enriched particles on a sucrose gradient (Batisse et al., 2009). They calculated the length of mRNAs packaged in particles of each fraction by performing Northern blot analysis using radioactively labelled poly dT probes. Combining these, they demonstrated that mRNPs are almost 11-fold compacted inside the cell (~15-30 nm length x ~5 nm width). They postulated that a 1 kb RNA which would be ~340 μm in length if stretched out, would form a ~30 μm particle when it is packaged as an RNP. Another study aimed to understand the packing of a specific long Balbiani ring mRNA from *Chironomus tentans* (Skoglund et al., 1983). This particular mRNA expressed in salivary gland nuclei, is 35-40 kb long which made the step-wise visualization of its packaging into mRNPs under EM possible. The authors followed the transcription of this mRNA and observed that it

underwent co-transcriptional packaging and compaction. As the mRNA was being transcribed, it formed a 19 nm wide rod. This started folding on itself at the distal end to form a 26 nm wide bulb at the top and 19 nm wide stalk. It was then released into the nucleoplasm where it collapsed to form a 50 nm wide globular structure. Finally, it is threaded through the nuclear pore as a rod-like structure. Both these studies demonstrated visually the extent of mRNA compaction inside the cells. However, to understand more general rules of mRNP packaging at a transcriptome-wide level, there is a need for development of high-throughput method.

Methods to capture higher order organization of polynucleotides

A fully stretched out human genome would measure ~ 2m in length. This along with its associated proteins needs to be packaged in sphere of ~20 μm diameter. To capture the 3D higher order structure of chromatin within the nucleus, a technique called Hi-C (Lieberman-Aiden et al., 2009) and its precursors were developed (de Wit and de Laat, 2012). Briefly, Hi-C involves cross-linking cells to freeze interactions, digesting DNA with a specific restriction enzyme that create breaks which can be then ligated to capture spatially proximal interactions. These interactions containing chimeric DNA pieces can then be purified and sequenced to understand how 3D organization of DNA inside cells. Thus, proximity ligations provide an established method to capture higher order structure of chromatin.

Proximity ligation methods developed for RNAs (Figure 1.3)

Cross-Linking and Sequencing of Hybrids (CLASH):

Proximity ligations were first used for RNA in a technique called CLASH (Cross-Linking and Sequencing of Hybrids; (Kudla et al., 2011)), albeit to capture inter-RNA interactions in yeast. Briefly, CLASH entailed crosslinking yeast cells followed by immunoprecipitation with the protein complex of interest (here snoRNA binding proteins; Nop1, Nop56, Nop58). The RNPs were then treated with a RNase to create single strand breaks which could then be ligated with T4 RNA ligase I. Associated proteins were then digested with proteinase K and RNAs were isolated. The hybrid reads thus obtained were sequenced using deep sequencing. A small fraction of the all the reads sequenced (< 1%) contained 'hybrid' reads where 2 different RNAs or parts of the same RNA were ligated together. This provided the information of what RNAs were close together in space inside the pulled-down complexes. Later they applied a similar approach in human cells to identify *in vivo* miRNA-mRNA interactions (Helwak et al., 2013). The major difference between Hi-C and CLASH was that CLASH enriched for specific RNA-protein complexes of interest using immuno-precipitation- Nop proteins that bind snoRNAs and Ago1 which binds miRNA and targets mRNAs. Thus, CLASH was used to capture the inter-RNA interactome of a specific RNA binding protein. It demonstrated that proximity ligations could be employed to capture *in vivo* inter-RNA interactions.

Along with the biochemical approach, they also developed a bioinformatic tool called “Hyb” to map the chimeric reads obtained from CLASH (Travis et al., 2014). Hyb used local alignment tools to map to a given reference and selected for reads with two non-contiguous alignments while discarding all contiguously aligned reads. After calling these “chimeric” reads, they folded them computationally to identify possible base-pairing interactions between the non-contiguous fragments within each read. Hyb tested multiple different local alignment tools including- BLAST, BLAST+, BLAT, pBLAT and Bowtie2 and found the most reliable results in the shortest time using Bowtie2. By default, Hyb mapped to the transcriptome, since mapping to the genome identified a large number of false positives as exon-exon junctions could not be distinguished from ligation junctions. For each read, Hyb identified multiple alignments and ordered them by mapping score. They paired the best alignment with another alignment that has either a gap or overlap of 4 nt. These pairs were then called as “chimeric” alignments for the read.

RNA hybrid and individual-nucleotide resolution ultraviolet crosslinking and immunoprecipitation (hiCLIP):

hiCLIP used a similar approach to CLASH in that it performed CLIP on a dsRNA-binding protein (here, Staufen1) after UV cross-linking the protein to its bound RNA (Sugimoto et al., 2015). However, one major difference between the

2 approaches was that hiCLIP ligated an adapter between the 2 spatially proximal RNA fragments. By doing this they could accurately identify interacting RNA regions. The hiCLIP protocol consisted of the following steps, first they UV cross-linked RNA-proteins by irradiating the cells at 254 nm, prepared whole cell extracts and treated these extracts with 2 concentrations (high and low) of RNase I. This was followed by IP with an antibody against Stau1 and end-repair using PNK to create 5′-P nd 3′-OH. Next, they ligated 2 pre-adenylated adapters (A and B) in equimolar quantities to the 3′ ends of IP-ed RNAs using T4 RNA ligase 2 truncated K227Q (it only ligates pre-adenylated donor RNAs to acceptor 3′-OH RNAs (Viollet et al., 2011)). Adapter B had a phosphate group that was then removed using T4 PNK and the 2nd RNA fragment was ligated to adapter B's 3′ end using T4 RNA ligase I, overnight at 16 °C. The RNP complexes were then denatured using urea to only select Stau1 associated RNAs. Following this, CLIP protocol was followed till purification of RNAs. Using hiCLIP they were able to show that Stau1 bound base-pairs were mostly intra-RNA and these interactions were depletion from coding regions of highly expressed mRNAs. To their surprise they observed long-range duplexes in 3′ UTRs of mRNAs.

To analyze reads obtained from sequencing these hybrid RNAs, they first identified reads with adapter B present in them with at least 17 nt on each right and left of the adapter. They then separated the left and right arms of these reads and mapped them separately to the desired reference using Bowtie. Thus, having

the adapter sequence in their reads simplified the mapping pipeline significantly compared to CLASH.

Later, three similar methods (see below) tried an unbiased approach to understand inter- and intra-RNA interactions mediated by base-pairs instead of identifying interactions limited to a single RNP complex.

LIGation of Interacting RNA followed by high-throughput sequencing (LIGR-seq):

To identify *in vivo* intra-RNA and inter-RNA base-pairs, LIGR-seq cross-linked RNA duplexes *in vivo* (Sharma et al., 2016), ligated their ends and then enriched them before sequencing. It used a psoralen derivative 4′-aminomethyltrioxsalen (AMT), that intercalates nucleotides and upon UV irradiation at 365 nm forms reversible adducts between nearby pyrimidine residues. They then made cell extracts, treated it with DNase I to remove DNA and depleted ribosomal RNAs (rRNAs) to reduce rRNA contamination. The RNAs were then treated with limited S1 endonuclease which cut single stranded RNA regions, to create smaller RNAs fragments without nicking the base-paired regions. Overhangs of cross-linked base pairs were then ligated using circRNA ligase. Unligated overhangs and single-stranded RNA fragments were digested away using 3′-5′ RNase R, thus enriching for only RNA fragments that were involved in base-pairing. The cross-links were reversed by irradiating at 254 nm and deep

sequencing libraries were prepared using these RNA fragments as well as a -AMT control. Using the data thus obtained, LIGR-seq was able to identify novel interactions between snoRNAs and mRNAs. Specifically, they demonstrated that SNORD83B downregulated the expression of its target mRNAs- CYTH, SRSF3, NOP14 and RPS5.

To analyze their data, they also developed a bioinformatics suite, “Aligater”, that uses bowtie2 in local alignment mode specifically with the option “-k 50”. When specified, Bowtie2 searches for N (here 50) alignments for the given read. They tested multiple different values for N (100, 250, 500, 1000), and for their dataset it did not improve chimera detection but made the program significantly slower. Then then processed all alignments for each read in the output bam file, to identify the best possible chimeric alignment arrangement. They joined all possible combination of alignments within a read minus a gap penalty (a penalty of 48 which corresponds to 6 matches of prefect mapping quality) and then identified the best possible “path” based on the alignment scores (LIGQ) of the resulting chimeric alignment.

Sequencing of psoralen cross-linked, ligated, and selected hybrids (SPLASH):

Similar to LIGR-seq, SPLASH also used a modified psoralen to capture *in vivo* intra- and inter-RNA interactions (Aw et al., 2016). However, instead of AMT,

they used biopsoralen which is a biotinylated form of psoralen that forms reversible cross-links in RNAs. Similar to AMT, biopsoralen can enter cells, intercalate between RNA base-pairs and cross-link pyrimidines upon irradiating with 365 nm UV light. They then made cell extracts, purified RNAs, fragmented and enriched for base-paired RNAs by immunoprecipitation with Streptavidin beads that bind to biotin. Proximity ligation was performed overnight using T4 RNA ligase I. The cross-links from ligated RNAs were reversed by irradiating at 254 nm and deep sequencing libraries were made using this sample. With this, they were able to identify novel intra- and inter-RNA base-pairing interactions. These included previously unknown snoRNA-rRNA interactions in humans and yeast and mRNA-mRNA interactions between mRNAs present in the same cellular compartments.

They used the following bioinformatics analysis pipeline- the sequenced reads from their paired-end libraries had overlaps between read 1 and 2 and were therefore merged together. These merged reads were then mapped to a transcriptome using BWA-MEM with minimum alignment length of 20 (default is 30). To identify chimeric reads, they scanned primary alignments in the bam file for a split alignment (SA) tag which lists other alignments in a chimeric alignment for a read. They discarded any SAs that were less than 50 nt apart to focus the analysis specifically on long range interactions. Further, in their validation of ligations using PDB structures, they observed junctions less than 50 nt to be always “true” since these would always be close in structure given their short distances and thus lead to higher false positives.

Psoralen analysis of RNA interactions and structures (PARIS):

The third approach to capture *in vivo* transcriptome wide base-pairing interactions, PARIS (Lu et al., 2016), also used a LIGR-seq like approach. They used AMT to cross-link base-pairs within RNAs by irradiation with UV 365 nm. RNAs extracted from cells underwent limited digestion with RNAse S1 (cuts ssRNA) and with proteinase K to remove cross-linked proteins. RNAs were then purified by TRiZol extraction and digested with ShortCut RNase III (dsRNase) to further fragment RNAs. To enrich for cross-linked dsRNAs, the sample was electrophoresed on a 2D gel (first, 12% native polyacrylamide gel followed by 20% urea-TBE denatured polyacrylamide gel). Proximity ligation using T4 RNA ligase I was performed on the dsRNAs that were purified from the second gel. After ligation, cross-links were reversed by irradiating at 254 nm and deep sequencing was performed to elucidate *in vivo* base-pairing interactions. PARIS was able to identify alternative secondary structures in multiple intra- and inter-RNA interactions, e.g., U4:U6 snRNA dimer, 3′ UTR of TUBB mRNA and lncRNAs (XIST and MALAT1). They also identified two long hairpins spanning upto 7 kb in the first and last exon of XIST.

PARIS used STAR with modified parameters to map their data to the desired reference. The STAR parameters used were,

STAR --runMode alignReads --genomeDir STAR_index

--readFilesIn fastq_file --outFileNamePrefix name_prefix

--outReadsUnmapped Fastq --outFilterMultimapNmax100

--outSAMattributes All --alignIntronMin 1 --scoreGapNoncan -4

--scoreGapATAC -4 --chimSegmentMin 15 --chimJunctionOverhangMin 15

These parameters allowed for lower penalty for gapped reads and permitted chimeric read alignment. STAR produced two output files, one for non-chimeric alignments and others with chimeric alignments. PARIS combined chimeric alignments and reads with large gaps to create a gapped-read file containing all putative duplexes. They created a pipeline that processed and grouped these reads into clusters called “duplex groups” representing a base-paired region.

While all four methods, CLASH, LIGR-seq, SPLASH and PARIS, used proximity ligations in conjunction with cross-linking to specifically probe for base-pairing interactions genome-wide, two other methods, RNA Proximity Ligation (Ramani et al., 2015) and Mapping RNA interactome in vivo (Nguyen et al., 2016), were designed to capture all interactions, base-pairing as well as protein mediated RNA folding, within an RNA.

RNA Proximity Ligation (RPL):

In RPL, proximity ligations were performed *in situ* by permeabilizing cells and treating them with T4 RNA ligase I under dilution conditions (Ramani et al.,

2015). To create ends for ligation, they either allowed endogenous RNases in yeast cells or used RNace-IT (mixture of RNase A1 and T1) for human cells to create nicks. All ligations were performed under native conditions without cross-linking to capture the more stable interactions. After overnight ligation, RNAs were purified using acid guanidinium-phenol and standard RNA-seq libraries were prepared. Unlike other techniques, RPL did not enrich specifically for chimeric RNA species. Further, without specifically selecting for base-pairs, they were able to capture interactions between proximal regions abundant RNAs (rRNAs, snoRNA snR86, U1 spliceosomal RNA snR19, RNA component of the Signal recognition particle SCR1) in yeast. Even though base-pairing interactions dominated the RPL dataset, it demonstrated that similar to that of chromatin, proximity ligations could be used to capture higher order RNA organization.

The paired-end reads sequenced were first merged using SeqPrep since some read pairs contained overlapping regions. Single reads thus obtained were mapped to the respective reference (yeast or human) using STAR with option for chimeric junction output. Parameters used for mapping were as follows-

--outSJfilterOverhangMin 6 6 6 6

--outSJfilterCountTotalMin 1 1 1 1

--outSJfilterDistToOtherSJmin 0 0 0 0

--alignIntronMin 10

--chimSegmentMin 15

--chimScoreJunctionNonGTAG 0

--chimJunctionOverhangMin 6

STAR generated output for chimeric alignments was analyzed using a custom set of scripts to first filter out known splice junctions and then parse chimeric junctions obtained most likely as a result of the ligase.

Mapping RNA interactome in vivo (MARIO):

MARIO aimed to capture *in vivo* transcriptome-wide intra- and inter-RNA interactions cross-linking and immobilizing RNA-protein complexes on beads, followed by ligation of a biotinylated linker in between the two RNA fragments that are spatially proximal (Nguyen et al., 2016). In MARIO, proteins that are directly in contact with RNA were crosslinked using UV irradiation at 254 nm to capture *in vivo* interactions. Cell extracts made from these cross-linked cells were then treated with RNase I to achieve fragment lengths of approximately 1,000-2,000 nt and washed under stringent conditions to remove non-specific protein-RNA interactions. RNA-associated proteins were then biotinylated using EZlink Iodoacetyl-PEG2-Biotin so the RNP complexes can be immobilized on streptavidin beads. Once immobilized, a biotinylated linker was ligated to 5′ ends of the bead-bound RNAs using T4 RNA ligase I. This is followed by ligating the 5′ of the linker

with the 3′ end of RNA that is spatially proximal under dilute conditions. Following the two-step ligation, RNAs were extracted using denaturing elution, associated proteins were digested using proteinase and RNAs were purified using PCIA extraction. This purified RNA was used for producing deep sequencing libraries upon rRNA depletion. MARIO was able to identify multiple intra- and inter-RNA interaction hot-spots that were mediated by base-pairing interactions.

MARIO suit of bioinformatics tools was able to identify proximally ligated RNA fragments using the linker that was ligated between the two RNA fragments in close proximity. MARIO uses local alignment (BLAST) to first classify reads based on the ligated RNAs fragments and linker (e.g., RNA1-linker-RNA2, RNA1-linker, linker-RNA2 or linker only). Once paired-end reads with the format RNA1-linker-RNA2 are identified, they map individual reads to the relevant genome using Bowtie or Bowtie2 in local alignment mode to identify which RNA fragments were spatially proximal and connected by a linker. In the final step, MARIO generated separate clusters of high coverage for each RNA regions. They then counted the interaction frequency for these RNA fragment clusters to identify interaction hot-spots.

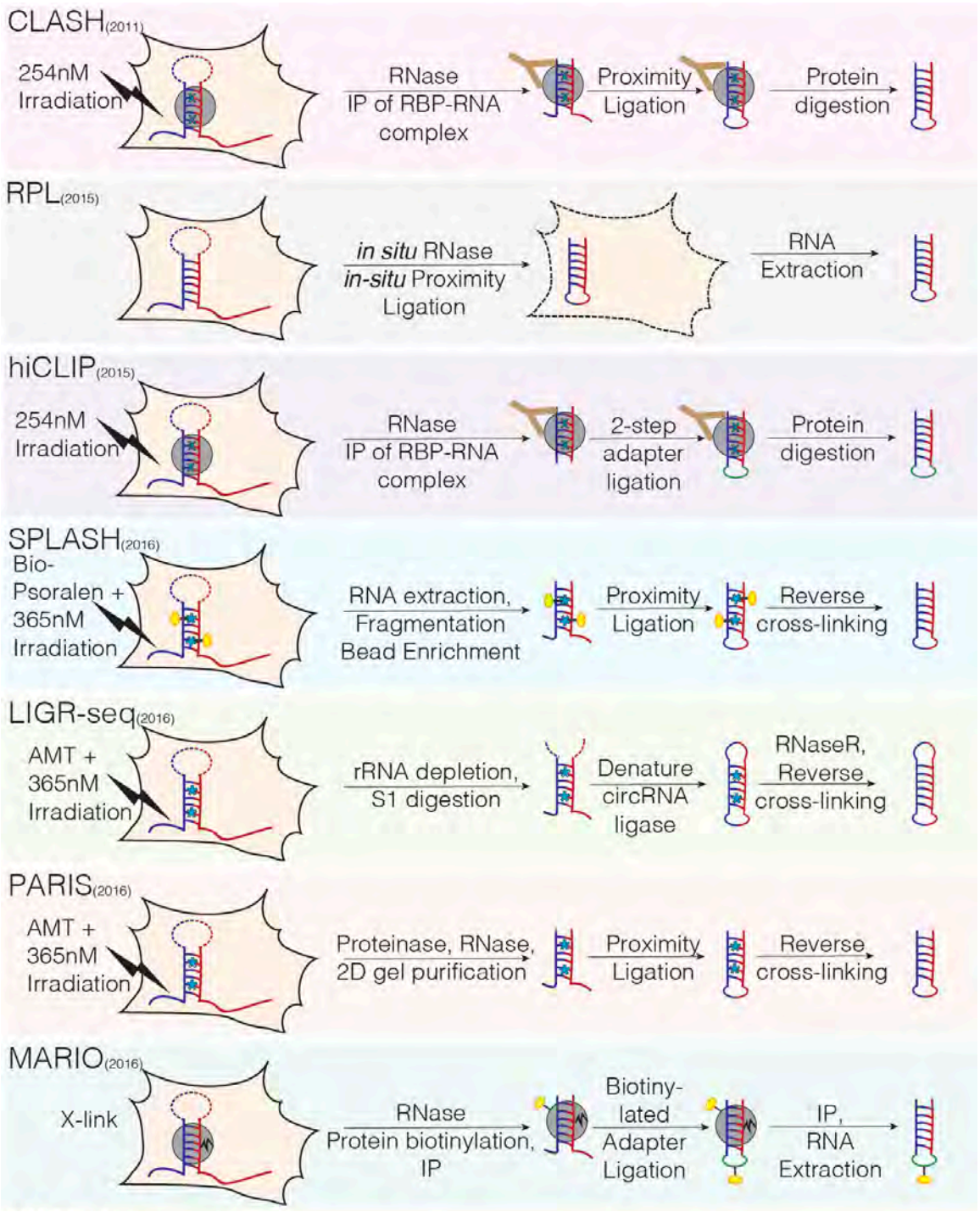

CLASH(2011)
254nM Irradiation
RNase IP of RBP-RNA complex
Proximity Ligation
Protein digestion
RPL(2015)
in situ RNase
in-situ Proximity Ligation
RNA Extraction
hiCLIP(2015)
254nM Irradiation
RNase IP of RBP-RNA complex
2-step adapter ligation
Protein digestion
SPLASH(2016)
Bio-Psoralen + 365nM Irradiation
RNA extraction, Fragmentation Bead Enrichment
Proximity Ligation
Reverse cross-linking
LIGR-seq(2016)
AMT + 365nM Irradiation
rRNA depletion, S1 digestion
Denature circRNA ligase
RNaseR, Reverse cross-linking
PARIS(2016)
AMT + 365nM Irradiation
Proteinase, RNase, 2D gel purification
Proximity Ligation
Reverse cross-linking
MARIO(2016)
X-link
RNase Protein biotinylation, IP
Biotiny-lated Adapter Ligation
IP, RNA Extraction

Figure 1.3. Schematic comparison of RNA proximity ligation methods

Overview of methods employing RNA proximity ligations. Lines in red and blue are spatially proximal RNA fragments coming from either the same transcript or two different transcripts. Gray discs: RBPs, blue asterisk: AMT intercalated within RNA, blue cross: UV crosslinked bases, black wiggly lines: protein-RNA crosslink, yellow-ovoid disc: biotin, brown “Y”-shape: antibodies, green line: oligonucleotide adapter, nude structure: cell, lightening shape: UV irradiation.

Secondary structural promiscuity in mRNAs

Overall, data from all RNA proximity ligation methods prove that proximity ligations can capture interactions between regions of RNAs that are close together in space. However, given that most of the RNA inside cells is non-coding (Palazzo and Lee, 2015), there is a need for mRNA enrichment in order to study higher order mRNA structures. Since many of the abundant non-coding RNAs are either involved in providing structural support (e.g., XIST) or part of an enzymatic machinery (e.g., rRNA, snRNA), they tend to have conserved structures (Rivas et al., 2001). Hence, interactions within ncRNAs are also conserved and therefore more likely to be captured even in datasets not biased towards base-pairing interactions.

However, we do not expect such specific contacts in higher-order mRNP structures. In fact, methods like DMS-seq (Rouskin et al., 2014) and icSHAPE (Spitale et al., 2015), were specifically designed to measure *in vivo* RNA secondary structure. They used chemical compounds like DMS (Rouskin et al., 2014) and NAI-N_3 (Spitale et al., 2015) that could easily enter cells and modify solvent accessible (i.e. single stranded) nucleotides while residues involved in base-pairing interactions are protected. These chemically modified residues (A and C for DMS-seq and all nucleotides for icSHAPE) block reverse transcription. Hence, by size selecting these blocked cDNA products and performing deep-sequencing on these, residues that were not involved in base-pairing could be identified. These studies used naked RNA that is heat denatured (95 °C) and re-folded as controls

to study the difference in secondary structure *in vivo* vs *in vitro*. Broadly, these secondary structure probing methods have demonstrated that ncRNAs like rRNAs (Rouskin et al., 2014) and XIST (Smola et al., 2016) are highly structured *in vivo* (Spitale et al., 2015) based on the reactivity of nucleotides. On the contrary, mRNAs demonstrated higher reactivity (and therefore more flexibility) *in vivo* compared to *in vitro* (Rouskin et al., 2014; Spitale et al., 2015). Thus, DMS-seq and icSHAPE data predicted that mRNAs are much more flexible in cells compared to naked RNAs folded *in vitro*. However, an important point to note is that these are ensemble experiments. Therefore, the apparent secondary structural variability may arise due to either mRNAs being less structured *in vivo* or that there are no specific points of interactions as in ncRNAs. So, each mRNA has a slightly different structure such that points of contacts are nearby but not the same. This would also give rise to more residues appearing accessible in ensemble experiments.

Conformational prediction for biopolymers

Though secondary structure probing methods provide information about the flexibility of the probed nucleotide chain, they impart very little information about the 3D arrangement or the conformation of the involved nucleotide chain. For example, secondary structural information does not allow us to predict if RNAs are

tightly packed or loosely packed inside an mRNP, or as to what shapes do mRNPs assume.

Since RNAs are biopolymers, principles of polymer physics can be applied to understand polymer arrangement inside a 3D structure. Indeed, polymer physics was able to predict the arrangement of DNA inside interphase chromosomes (Lieberman-Aiden et al., 2009) and mitotic chromosomes (Gibcus et al., 2018; Naumova et al., 2013) using data obtained from proximity ligations for DNA. Since each locus has an ensemble of interactions, a contact probability distribution can be obtained for that locus with all other loci in the genome. These interaction probabilities decay with increasing distance. Any locus will have a higher chance of interacting with another locus that is spatially proximal rather than one that is distal (Fudenberg and Mirny, 2012). Plotting these interaction probabilities against the linear distance between the corresponding loci can provide insights into the polymer conformation. In polymer physics there are 3 basic polymer types (Fudenberg and Mirny, 2012)- random coil (random walk), the swollen coil (self-avoiding walk), and the equilibrium globular state. A brief and simplified explanation of each polymer state is as follows-

Random coil: A polymer adopts an ensemble of conformations called random walk if the steric repulsion between the monomer of a polymer are balanced by its interactions with the solvent and its topological constraints are disregarded. In this conformation, the chain is unconstrained in 3D and the polymer overall is loosely

packed. Therefore, the contact probability between monomers decays rapidly, P(s) ~ $s^{3/2}$, where s is the distance between the monomers along the polymer.

Swollen coil: In this type of polymer, the steric repulsion between monomers is not negligible. This means that the monomers try to avoid each other and therefore the polymer conformations are even more loosely packed compared to random coil. The contact probability between monomers decays more rapidly than random coil, P(s) ~ $s^{0.6}$.

Equilibrium globule: A polymer assumes this ensemble of conformations if there is attraction between monomer, repulsion from solvent and the polymer is confined to a small volume. Equilibrium globules are highly compact and space filling. At short distances, they behave like random coils, therefore contact probability is given by P(s) ~ $s^{-3/2}$. However, at longer lengths they are constrained by the confinement, the polymer "bounces" back leading to increasing in interaction frequency and the *P(s)* plateaus after the initial drop.

In the case of interphase chromatin, it was predicted to be packaged as fractal globule at the scale of several megabases (Lieberman-Aiden et al., 2009). The fractal globule is a non-equilibrium ensemble of conformations. Like equilibrium globules, the polymer is highly compacted, however, the polymer itself is unknotted. The polymer crumples into a series of small globules which fold on themselves until the whole polymer forms a single compact globule. The fractal globules are characterized by $P(s) \sim s^{-1}$ *instead of* $P(s) \sim s^{-3/2}$ as seen for

equilibrium globules. It was hypothesized that the unknotted nature of fractal globules assists in chromosome unfolding and refolding required for different activities like gene expression and repression.

Similar analysis using contact probabilities through different stages of mitosis, demonstrated that the mitotic chromosome is arranged as a rod-like structure (Gibcus et al., 2018). Inside this, the chromosomes adopt a spiral staircase-like structure that is organized along alpha helical protein axis formed by Condensin I and II.

Together, these studies demonstrate that polymer physics theories can be applied to biopolymers to understand overall shape as well as organizational principles of these molecules.

Concluding remarks

In conclusion, very little is known about the structure/conformation of information encoding molecules, mRNAs and their associated proteins. Given the amount of cellular abundance of other ncRNAs like rRNAs and snRNAs which constitute more than 95% of the cellular RNA there is a need for a novel approach to specifically capture organization principles of mRNPs. In addition, the approach needs to capture not only base-pairing interactions but also other interactions (like those mediated proteins) to re-shape RNAs and bring together its different parts. This thesis will introduce and discuss the development of one such approach,

RIPPLiT (RNA ImmunoPrecipitation and Proximity Ligation in Tandem), a transcriptome-wide method for probing the 3D conformations of RNAs. Further, there are no established bioinformatics tools to analyze the chimeric dataset obtained as a result of employing such a method. Hence, we also develop a novel suite, ChimeraTie, for processing, manipulating, analyzing and visualizing chimeric data obtained from RIPPLiT. I will also include a step-wise manual to perform RIPPLiT experiments as well performing the bioinformatics analysis for those datasets using ChimeraTie. Furthermore, this data and the application of polymer physics rules, will allow for the elucidation of some of the fundamental rules for mRNP packaging and organization.

Chapter II

Understanding the higher order structure of pre-translational mRNPs

ABSTRACT

Compared to noncoding RNAs (ncRNAs) such as rRNAs and ribozymes, for which high resolution structures abound, little is known about the tertiary structures of mRNAs. In eukaryotic cells, newly made mRNAs are packaged with proteins in highly compacted mRNPs, but the manner of this mRNA compaction is unknown. Here we developed and implemented RIPPLiT (RNA ImmunoPrecipitation and Proximity Ligation in Tandem), a transcriptome-wide method for probing the 3D conformations of RNAs stably-associated with defined proteins, in this case exon junction complex (EJC) core factors. EJCs multimerize with other mRNP components to form megadalton sized complexes that protect large swaths of newly synthesized mRNAs from endonuclease digestion. Unlike ncRNPs wherein strong locus-specific structures predominate, mRNPs behave more like flexible polymers. Polymer analysis of proximity ligation data for hundreds of mRNA species demonstrates that nascent and pre-translational mammalian mRNAs are compacted by their associated proteins into linear rod-like structures.

INTRODUCTION

Once synthesized, messenger RNA particles (mRNPs) must explore the nuclear and cytoplasmic compartments to arrive at their final subcellular destinations. Such travel necessitates packaging in a manner that prevents RNA tangling, shearing and premature degradation. To date, however, the rules governing RNA polymerase II (Pol II) transcript packaging remain largely undefined. Assays such as icSHAPE and DMS-seq that detect nucleotide accessibility suggest that mRNAs are generally flexible and unstructured, even more so *in vivo* than *in vitro* (Rouskin et al., 2014; Spitale et al., 2015). Such studies, however, provide no information regarding the conformational properties of the RNA polymer inside the mRNP, its degree of compaction or its overall shape. Conformationally, the RNA could fold as a simple random coil, an equilibrium globule, a fractal globule, or some other polymer arrangement, and this unstructured arrangement could be either loosely packed or highly compacted.

The only available data about overall mRNP shape in cells come from electron microscopy (EM) studies where images of purified polyA+ transcripts from budding yeast (in which mRNAs average ~1,250 nt (Miura et al., 2008)) revealed rod-like structures of differing lengths but nearly constant width (~5 nm) (Batisse et al., 2009). At the opposite extreme, *in situ* EM images of giant Balbiani ring mRNAs (35,000 to 40,000 nt) in *Chironomus tentans* salivary gland nuclei also showed rod-like structures (~10 nm wide) that collapse into 19 nm stalks during transcription and then 50 nm globular structures upon chromatin release (Skoglund

et al., 1983). For both yeast and Balbiani ring mRNPs, these measured dimensions necessitate substantial RNA condensation relative to a simple linear structure, with the estimated compaction being ~11-fold for yeast mRNAs and ~200-fold for Balbiani ring mRNAs. But how this is accomplished and what general principles guide mRNP 3D organization are currently unknown. Further, it is unknown whether the globular structures adopted by Balbiani ring mRNPs are unique to these exceptionally long transcripts. That is, are mammalian mRNPs (containing >3,900 nt mRNAs on average; NCBI *Homo sapiens* Annotation Release 108) more rod-like or more globular?

Pre-translational mRNPs contain tightly-bound proteins that accompany the mRNA from the nucleus to the cytoplasm. Chief among these are exon junction complexes (EJCs), composed of three core factors: eIF4AIII, Magoh and Y14. EJCs are assembled from their component parts upstream of exon junctions by the spliceosome during the process of intron excision. Once assembled they remain in place, accompanying the mRNA to the cytoplasm where they are removed by the first round of translation (Dostie and Dreyfuss, 2002; Lejeune et al., 2002). The mechanism of their sequence-independent deposition essentially locks EJCs onto the RNA until they are unlocked by a factor associated with elongating ribosomes (Gehring et al., 2009). Once disassembled, the individual proteins are rapidly reimported into the nucleus explaining their strong nuclear localization (Bono et al., 2010; Shibuya et al., 2004).

We previously showed that endogenous EJCs interact both with one another and with other tightly bound mRNP components (e.g., peripheral EJC proteins, serine/arginine (SR)-rich proteins and SR-like proteins) through short and long-range interactions to form RNase-resistant, megadalton-sized complexes containing 30-150 nt protected fragments of spliced mRNAs (Singh et al., 2012). Incredibly, even in the absence of any crosslinking agent, these higher order structures are stable to stringent double IPs and nuclease treatments designed for footprinting assays, and have molecular weights exceeding 2 MDa (for comparison, the large ribosomal subunit has a molecular weight of 1.7 MDa). Thus, EJCs and their associated proteins form a large and stable structural core that packages and protects newly made mRNAs. Here we used RNA Proximity Ligation (Kudla et al., 2011; Ramani et al., 2015) to investigate the higher order structure, polymer compaction and RNA folding principles within this core.

RESULTS

RIPPLiT: A method to capture higher order RNA structure in RNPs

We previously developed RNA Protein Immunoprecipation in Tandem (RIPiT) (Singh et al., 2012; Singh et al., 2014) to enable purification of RNP complexes containing specified protein pairs. EJC-containing RNPs can be selectively enriched by first affinity-enriching for one EJC core protein (e.g., FLAG-tagged Magoh), and then immunopurifying a second (e.g., eIF4AIII). Inclusion of a

nuclease digestion step between the two affinity steps enables the production of footprints. Here we modified this protocol to include a proximity ligation step after nuclease digestion (**Figure 2.1A**). Briefly, we immunoprecitated FLAG-tagged EJC-containing particles and fragmented the bound RNA via limited RNase T1 digestion to generate a fragment distribution ranging from 30 to >500 nt (**Figure 2.1B**). Following conversion of RNA ends to 5′-P and 3′-OH groups (necessary for ligation) during the second immunoaffinity step (**Figure 2.1A**), T4 RNA Ligase I (Rnl I) addition produced a distinct shift toward larger sized RNA fragments (**Figure 2.1B**). Protection from phosphatase removal of 5′-^{32}P labels incorporated during the phosphorylation step confirmed that these longer fragments were bona fide ligation products (**Figure 2.2)**. We call this new method for mapping higher ordered structures within double affinity-purified RNPs RNA Protein Immunoprecipation and Proximity Ligation in Tandem (RIPPLiT).

For RNP structural analysis, we performed RIPPLiT on three independent biological replicate whole cell lysates from HEK293 cells expressing FLAG-tagged Magoh. Immediately before the ligation step, each replicate was divided in half, with one half receiving ligase (+ ligase) and the other not (- ligase). All libraries were size selected for ~200-550 nt inserts. Preliminary data analysis by Sanger sequencing revealed that the + ligase libraries contained a high fraction of chimeric reads (i.e., reads made up of concatenated fragments mapping to different locations on one or more RNA species) (**Figure 2.1C**), whereas the - ligase libraries did not. Paired-end sequencing (150 nt) on the Illumina NextSeq platform

yielded 23 to 49 million merged reads per library (**Table 2.1**). To facilitate chimeric read mapping, we developed ChimeraTie, a bioinformatics tool that employs the local alignment mode of Bowtie2 to iteratively map all fragments within a single read (**Figure 2.1D**). Pair-wise chimeric junctions are visualized as two-dimensional heatmaps, where color intensities correspond to junction frequencies. Aggregating counts into appropriate length bins (e.g., 100 bins each corresponding to 1% total transcript length) assists in data visualization across long RNAs (**Figure 2.1D.3**).

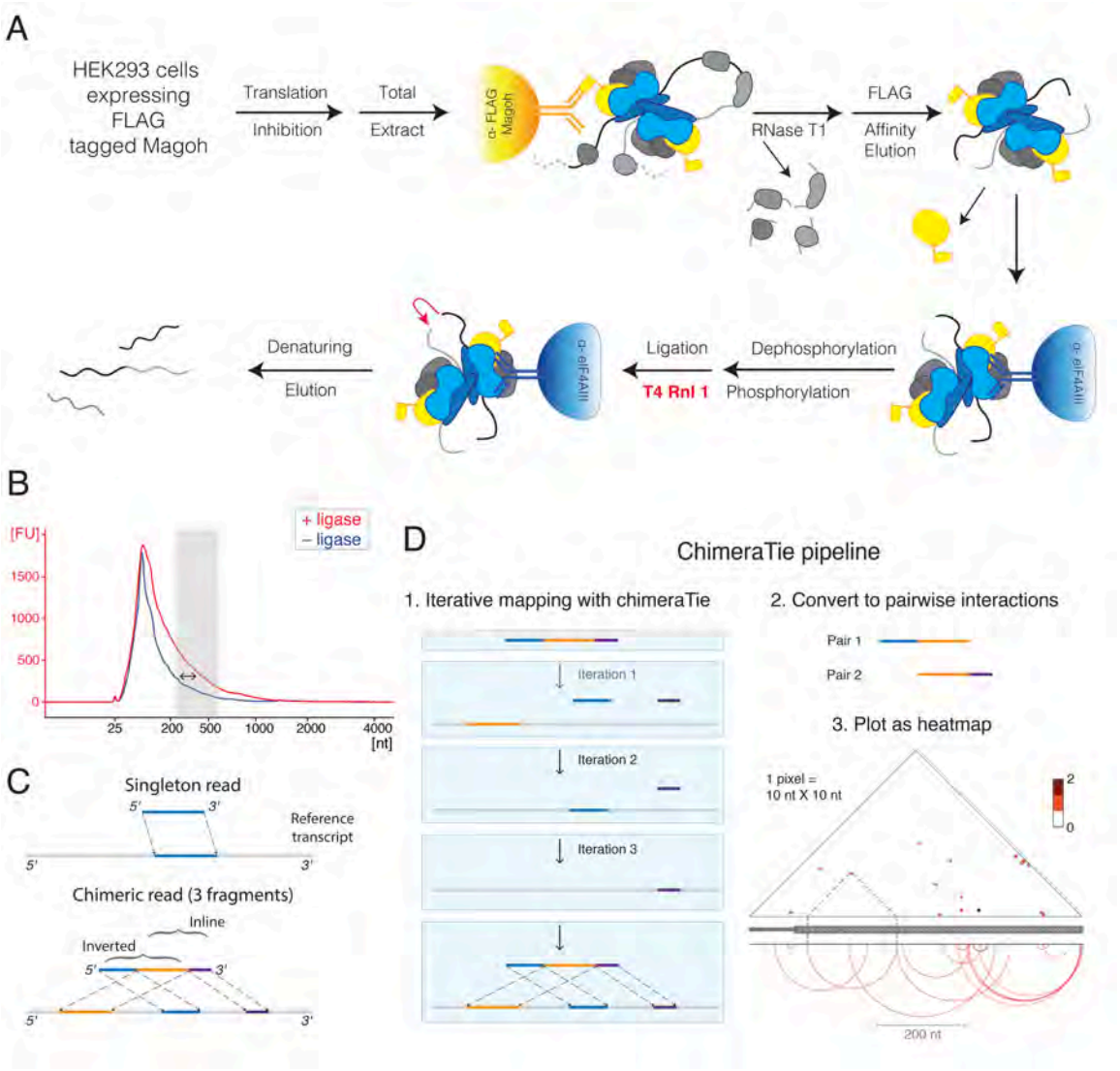


Figure 2.1. Overview of RIPPLiT and ChimeraTie

(A) Schematic for EJC RIPPLiT. Solid yellow and blue objects, EJC core proteins; gradient yellow and blue objects, antibody-conjugated beads; gray objects, non-EJC proteins; black and grey lines, RNA; red arrow, proximity ligation event.

(B) Bioanalyzer trace showing length distribution of RNAs obtained from EJC RIPPLiT (replicate 1). Double-headed arrow, length distribution shift due to addition of ligase; gray box, sizes selected for sequencing.

(C) Types of reads, fragments and chimeric junctions in RIPPLiT libraries, and their relationships to reference transcripts. Dotted lines indicate the fragment ends involved in chimeric junctions.

(D) Schematic of ChimeraTie pipeline used to iteratively map fragments, extract pairwise chimeric junctions and then visualize junctions as a heatmap. Data shown are chimeric junctions (replicate 1 only) within the first 767 nts of PRPF8 mRNA. Thick gray line with arrowheads, coding exons; thinner section; 5' UTR. Arcs show individual chimeric junctions at nt resolution. Heatmap indicates number of junctions within each 10 x 10 nt pixel, with dotted lines indicating the heatmap position of one chimeric junction.

Table 2.1. Number of PEAR-merged reads and uniquely mapping fragments for - and + ligase RIPPLiT libraries

	Reads	Uniquely Mapping Fragments		
	Total	**rRNA**	**snRNA**	**Transcriptome**
Rep1 + ligase	48,977,271	5,703,148	265,579	15,548,957
Rep2 + ligase	39,191,183	4,368,298	102,384	11,219,829
Rep3 + ligase	49,670,503	4,706,263	301,330	12,181,352
Rep1 - ligase	23,232,051	2,059,236	123,337	5,652,808
Rep2 - ligase	28,498,115	2,398,647	66,130	6,378,783
Rep3 - ligase	25,328,887	2,086,898	75,841	5,631,999

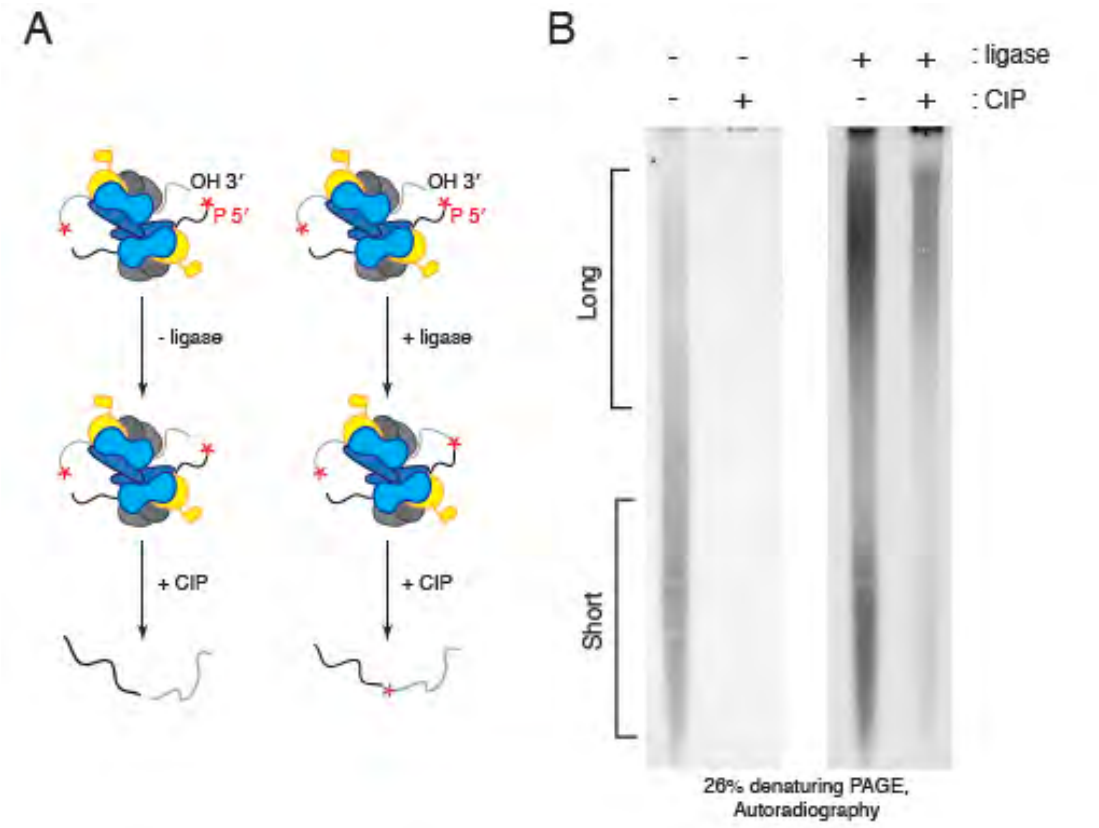


Figure 2.2. Shift in RNA size is due to ligation

(A) Schematic for 5′ end protection assay. Orange structures are EJCs; red and blue lines are RNA fragments. Red asterisk indicates radioactive phosphates.

(B) Urea PAGE for 5′ end protection assay. Black left brackets indicate short and long EJC footprints.

Confirming our purification of EJC-associated RNAs, transcripts from multi-exon genes were enriched over those from single exon genes in RIPPLiT libraries compared to RNA-seq (**Figure 2.3A, 2.4A**). In the + ligase libraries, spliced transcripts also had more chimeric junctions than single exon transcripts (**Figure 2.3B,C, 2.4B**). Both inline and inverted chimeric junctions were well represented and distributed across a wide range of intervening nucleotide distances (aka, spans) in + ligase libraries (**Figure 2.3D, cumulative distribution**). In contrast, - ligase chimeric junctions were predominantly inline (**Figure 2.3D, inset Table**) and these inline junctions were strongly skewed toward shorter spans (**Figure 2.3D, cumulative distribution**). These - ligase inline junction features are characteristic of mapping artifacts wherein a long single fragment is mistakenly mapped as two or more shorter inline fragments separated by short gaps. To minimize this issue, we optimized Bowtie2's read and reference gap penalties to disfavor short fragment junction gaps, and we used only uniquely mapping fragments (i.e., having no secondary alignment) with fewer than 3 mismatches/indels for all analyses. Nonetheless, since the relative weighting of a gap penalty decreases as alignment length increases in Bowtie2, we were unable to completely eliminate all false-positive, inline chimeric junctions. With these filters in place, intramolecular chimeric junctions were 3.5 to 8-fold higher in + than - ligase libraries (**Figure 2.3D, inset Table; Table 2.2A**).

Strong enrichment for intramolecular chimeric junctions in + over - ligase libraries was readily apparent upon examination of individual transcripts (**Figure**

2.4C). Further, despite differences in ligation efficiency between biological replicates (with + ligase chimeric junction counts in replicate 1 > replicate 2 > replicate 3), the number of chimeric junctions per transcript was highly correlated ($r > 0.96$) across all three replicates (**Figure 2.3E**). Thus, intramolecular mRNA ligations were abundantly represented in our datasets and exhibited high biological reproducibility.

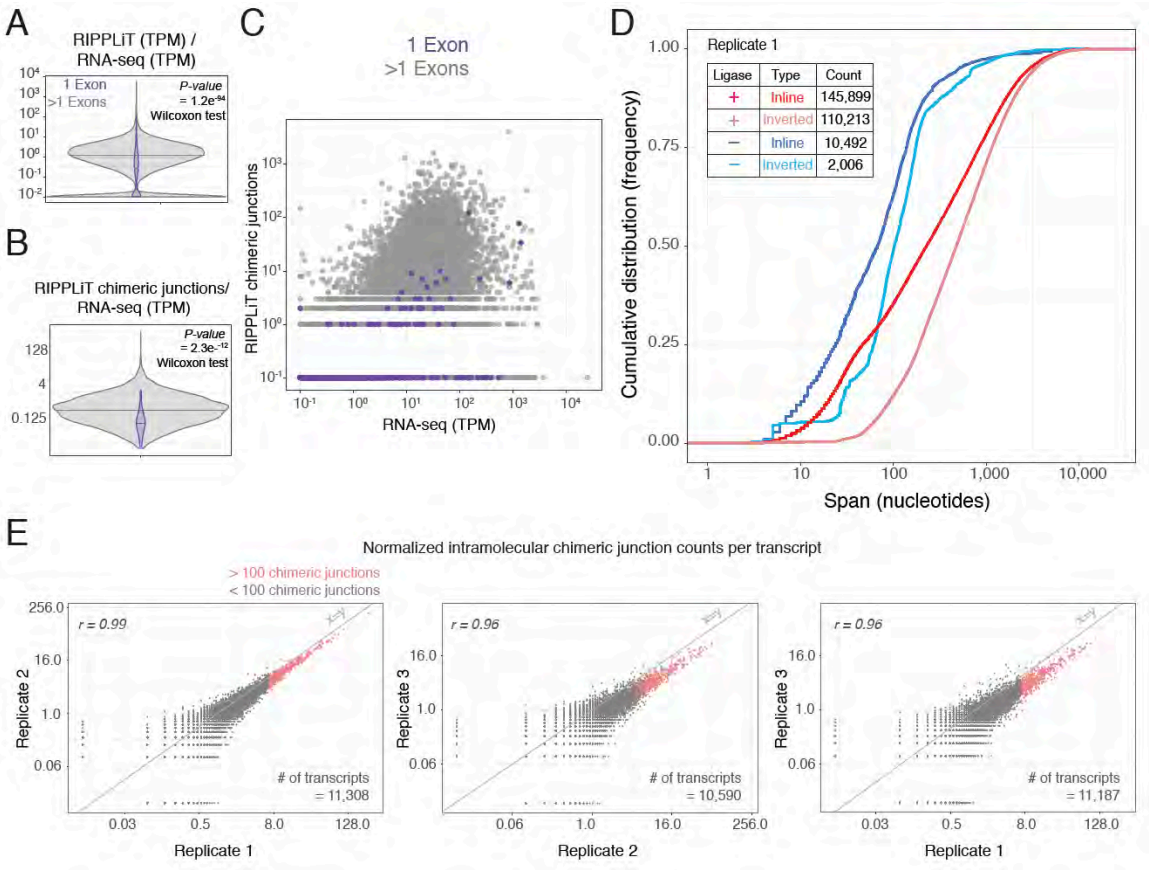


Figure 2.3. RIPPLiT captures intramolecular ligations in EJC-associated RNAs with high reproducibility

(A,B) Overlaid violin plots for RIPPLiT (transcripts per million: TPM; replicate 1 - ligase) over RNA-seq (TPM; (Ge et al., 2016)) and (B) chimeric junction count (replicate 1 + ligase) over RNA-seq (TPM) for genes with one (purple) or more than one (gray) exon. Gray and purple horizontal lines: medians. Genes not detected by RNA-seq were omitted from (B). TPM values were obtained by mapping RIPPLiT - ligase and RNA-seq libraries to human genome GRCH37 with RSEM; chimeric junction counts were obtained by mapping RIPPLiT - and + ligase libraries to the HEK293 transcriptome with ChimeraTie.

(C) Scatter plot comparing RIPPLiT chimeric junction counts to RNA-seq TPM.

(D) Cumulative frequency distributions of inline and inverted chimeric junction spans for replicate 1 - and + ligase. Inset table: Raw junction counts.

(E) Scatter plots comparing normalized intramolecular chimeric junction counts per transcript in + ligase libraries among biological replicates. Diagonal line: x=y. Red dots: set of transcripts used for scaling plots in **Figure 2.13**.

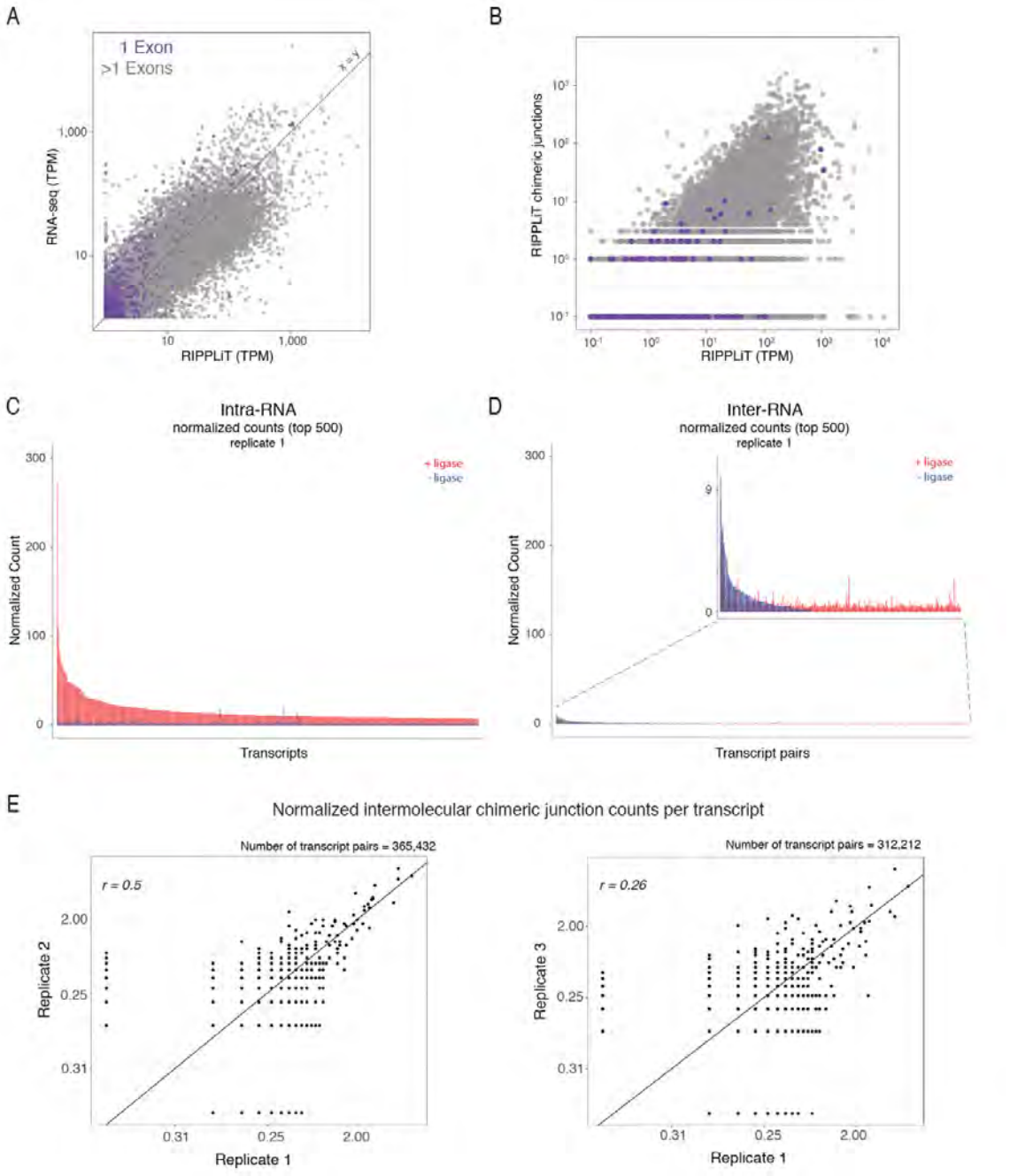


Figure 2.4. Chimeric junctions are less abundant on single exon genes and between transcripts (inter-RNA) than within individual transcripts (intra-RNA) from multiexon genes

(A) Scatter plot comparing fragment coverage of RNA-seq (transcripts per million: TPM; Ge et al., 2016) and RIPPLiT (replicate 1 - ligase; TPM) for genes with one (purple) or more than one (gray) exons. Grey line: x = y.

(B) Scatter plot comparing RIPPLiT chimeric junction count for genes with one (purple) or more than one (gray) exon in RIPPLiT (replicate 1 + ligase) to RNA-seq (TPM).

(C) Red: Histogram showing normalized intramolecular (intra-RNA) chimeric junction counts for top 500 Pol II transcripts in replicate 1 + ligase library. Blue: Overlaid histogram showing chimeric junction counts for the same transcripts in replicate 1 - ligase library. Transcripts are ordered by their chimeric junction counts in the + ligase library.

(D) Red: Histogram showing normalized intermolecular (inter-RNA) chimeric junction counts for top 500 Pol II transcript pairs in replicate 1 + ligase library. Blue: Overlaid histogram showing chimeric junction counts for the same transcript pairs in replicate 1 - ligase library. Transcripts are ordered by their chimeric junction counts in the - ligase library.

(E) Scatter plots comparing normalized intermolecular chimeric junction counts per transcript pair between + ligase biological replicates. Black line: x = y.

Table 2.2. Mapping statistics for human ribosomal RNAs and transcriptome

A

TRANSCRIPTOME	Rep1 plus		Rep1 minus		Rep2 plus		Rep2 minus		Rep3 plus		Rep3 minus	
# of uniquely mapping sense fragments	15,060,203		5,459,711		10862632		6164826		11,774,766		5,449,595	
Plus v/s Minus ratio	2.8				1.8				2.2			
Read distribution	#_fragments	Count	#_fragments	Count	#_fragments	Count	#_fragments	Count	#_fragments	Count	#_fragments	Count
	1	13,294,420	1	5,538,415	1	9,909,645	1	6,242,740	1	11,454,233	1	5,510,204
	2	1,027,028	2	53,812	2	605,693	2	64,473	2	345,477	2	57,793
	3	62,404	3	2,026	3	30,929	3	2,175	3	11,368	3	1,913
	4	3,156	4	154	4	1,436	4	129	4	485	4	110
	5	123	5	15	5	51	5	10	5	23	5	6
	6	5	6		6	2	6	1	6	1		
Total # reads containing mapped fragments	14,387,136		5,594,422		10,547,754		6,309,528		11,811,587		5,570,026	
# of chimeric reads	1,092,716		56,007		638,111		66,788		357,354		59,822	
Chimeric reads fold change	7.1				5.4				2.8			
Percent chimeric reads	7.6		1.0		6.0		1.1		3.0		1.1	
Intra-RNA "Direct" junction pairs	253,995		12,048		147,999		15,259		95,995		13,623	
Normalized Intra-RNA "Direct" junctions fold change	7.6				5.5				3.3			

B

rRNA	Rep1 plus		Rep1 minus		Rep2 plus		Rep2 minus		Rep3 plus		Rep3 minus	
# of uniquely mapping sense fragments		5,697,722		2,057,723		4,362,829		2,396,346		4,701,941		2,085,457
Plus v/s Minus ratio		2.8				1.8				2.3		
Read distribution	**#_frag-ments**	**Count**	**#_frag-ments**	**Count**	**#_frag-ments**	**Count**	**#_frag-ments**	**Count**	**#_frag-ments**	**Count**	**#_frag-ments**	**Count**
	1	5,286,673	1	2,049,405	1	4,100,312	1	2,388,130	1	4,576,606	1	2,080,100
	2	198,370	2	4890	2	128,413	2	5,212	2	63,506	2	3,378
	3	6,278	3	17	3	3,582	3	31	3	867	3	1,913
	4	219			4	101			4	11	4	14
	5	5			5	2						
Total # reads containing mapped fragments		5,491,545		2,054,312		4,232,410		2,393,373		4,640,990		2,083,492
# of chimeric reads		204,872		4,907		132,098		5,243		64,384		3,392
Chimeric reads fold change		15.1				13.8				8.4		
Percent chimeric reads		3.7		0.2		3.1		0..2		1.4		0.2
Intra-RNA "Direct" junction pairs		112,087		1,578		72,247		1,361		40,265		1,018
Normalized Intra-RNA "Direct" junctions fold change		25.7				29.2				17.5		
Reads Mapping to each reference	**Ref**	**Count**	**Ref**	**Count**	**Ref**	**Count**	**Ref**	**Count**	**Ref**	**Count**	**Ref**	**Count**
	18S	3,096,819	18S	891,462	18S	2,547,177	18S	1,259,819	18S	2,769,474	18S	1,197,278
	28S	2,287,536	28S	991,402	28S	1,601,630	28S	995,569	28S	1,466,603	28S	698,637
	5.8S	315,593	5.8S	175,672	5.8S	217,717	5.8S	142,834	5.8S	466,813	5.8S	190,501
	5S	3,200	5S	700	5S	1,774	5S	425	5S	3,373	5S	482
Inter-RNA junctions												
28S :: 18S		19,060		70		11,813		88		5,196		42
28S :: 5.8S		5,507		10		3,163		5		2,913		2
28S :: 5S		11		1		8		0		7		0
18S :: 5.8S		972		0		507		1		295		2
18S :: 5S		42		0		21		0		25		0
5.8S :: 5S		0		0		0		0		1		0

RIPPLiT faithfully captures ribosomal subunit 3D structure

Due to the high cellular abundance of ribosomes, ~27% of uniquely mapping RIPPLiT fragments mapped to rRNAs, with 18S rRNA exhibiting the highest fragment density (**Table 2.1**, **Table 2.2B**). Because the folded structure of 18S rRNA within the small ribosomal subunit is well known, we could use these data to test the validity of our method. Whereas + and - ligase libraries exhibited similar fragment coverage across the entire 18S rRNA (**Figure 2.5A**, tracks along sides of heatmap), chimeric junctions were 18- to 26-fold more abundant in + ligase than in - ligase libraries (inset numbers in **Figure 2.5A,B, Table 2.2B**). This difference is readily apparent in the + and - ligase chimeric junction heatmaps (**Figure 2.5A**). Further, both inline and inverted chimeric junctions occurred with near equal frequency in the + ligase libraries (**Figure 2.5B**), and they produced a strong locus-specific heatmap pattern that was highly reproducible across all three biological replicates (**Figure 2.5A**).

Since all rRNA molecules adopt the same 3D structure, 18S chimeric junctions should be most prevalent at locations that are both accessible to nuclease digestion and close enough in Euclidean (3D) space for subsequent ligation. To assess whether our observed junction pattern was consistent with native 18S rRNA folding within the ribosome, we use a high-resolution human 80S ribosome cryo-EM structure (**Figure 2.5C**; (Khatter et al., 2015)) to calculate mean Euclidean phosphate-phosphate distances for every 5 nt bin. Overlaying the + ligase chimeric junction frequency and mean 3D distance heatmaps (**Figure 2.5D**)

revealed preferential association of chimeric junctions with 3D-proximal regions (yellow) compared to 3D-distal regions (purple). Because ligations require both accessibility and proximity, not all 3D-proximal regions were enriched for chimeric junctions. Nonetheless, a plot of mean 3D-distance vs chimeric junction frequency shows the expected decay in junction frequency as Euclidean distance increases (**Figure 2.5E**). This same relationship was observed for chimeric junctions mapping within 28S rRNA (**Figure 2.6**).

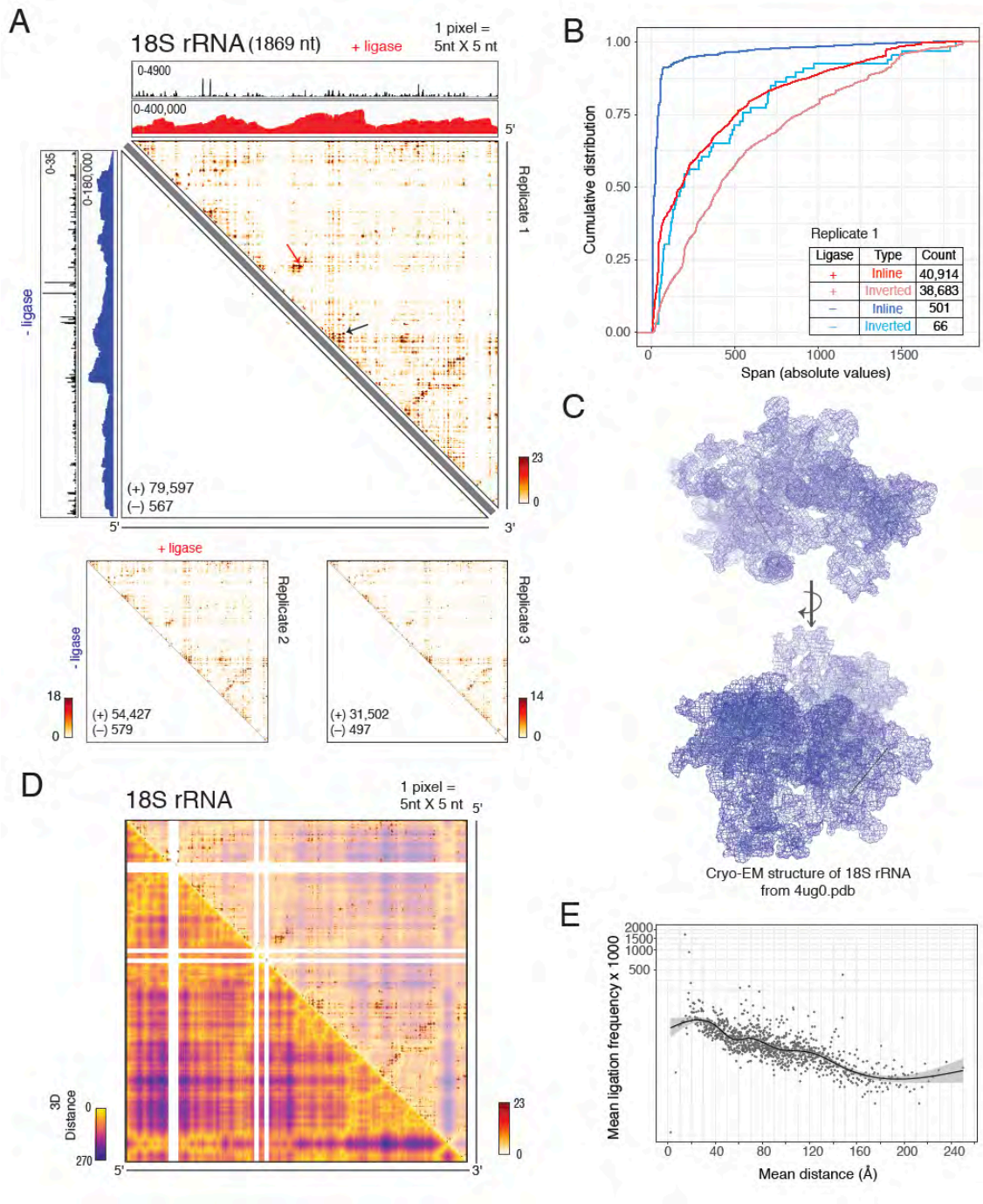


Figure 2.5. Ligations in 18S RNA occur between 3D-proximal regions

(A) Chimeric junction heatmaps for human 18S rRNA - (lower left) and + ligase (upper right). Color scales, number of junctions per 5x5 nt bin; numbers in lower

left corners, number of chimeric junctions obtained per indicated library. Replicate 1 coverage tracks show fragment distributions (red and blue) across the entire transcript and chimeric junction frequencies (black) at individual nucleotides. Note large scale differences for - and + ligase chimeric junction tracks.

(B) Cumulative frequency distributions of inline and inverted chimeric junction spans on 18S rRNA for - and + ligase libraries (replicate 1). Inset table: Number of inline and inverted chimeric junctions for each library.

(C) Structure of 18S rRNA (4UG0 (Khatter et al., 2015)) showing the two chimeric junctions marked with red and black arrows in (A) mapped onto the structure as lines.

(D) Heatmaps for mean Euclidean phosphate-phosphate distances (bottom) overlaid with chimeric junctions (top). White areas: regions absent from structure.

(E) Scatter plot showing mean ligation frequency in replicate 1 as a function of mean Euclidian distance for 1,000 bins each containing 79-80 chimeric junctions. Black line shows smoothing (GAM: generalized additive model) with grey area displaying confidence interval (0.95) around smoothing.

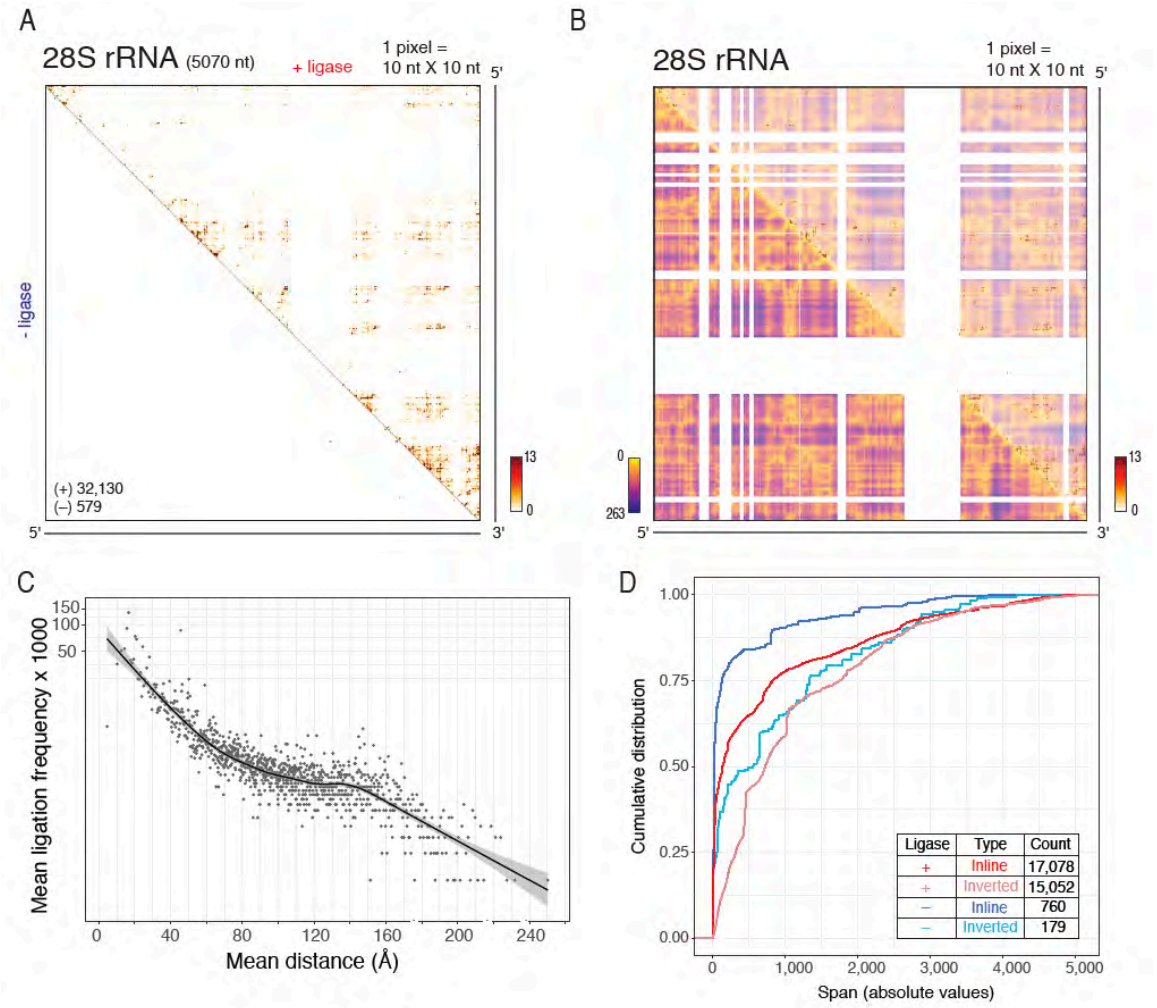


Figure 2.6. Ligations in 28S rRNA occur between 3D-proximal regions

(A) Same as Figure 2.5A, but for 28S rRNA.

(B) Same as Figure 2.5D, but for 28S rRNA.

(C) Same as Figure 2.5E, but for 28S rRNA.

(D) Same as Figure 2.5B, but for 28S rRNA.

RIPPLiT also captures inter-RNA proximities

In addition to intramolecular chimeric junctions, we also observed thousands of junctions for which the individual fragments mapped to different rRNAs (**Table 2.2B)**. These intermolecular junctions were 81 to 114-fold more prevalent in + ligase than - ligase libraries (**Table 2.2B**), suggesting that the vast majority of the + ligase junctions were not mapping artifacts. Mapping these onto the 80S ribosome structure resulted in a similar Euclidean distance versus junction frequency distribution as the intramolecular junctions (**Figure 2.7A**). In addition, most intermolecular junctions between the 5.8S and 28S rRNA occur near the 5′ end of the 28S rRNA (**Figure 2.7B**), the known 5.8S-28S interaction domain (Walker and Pace, 1983) (**Figure 2.7C**). Thus, RIPPLiT accurately reports 3D structural information in native RNPs, and can capture both intra- and intermolecular proximities within stable multi-RNA complexes with a high signal to noise.

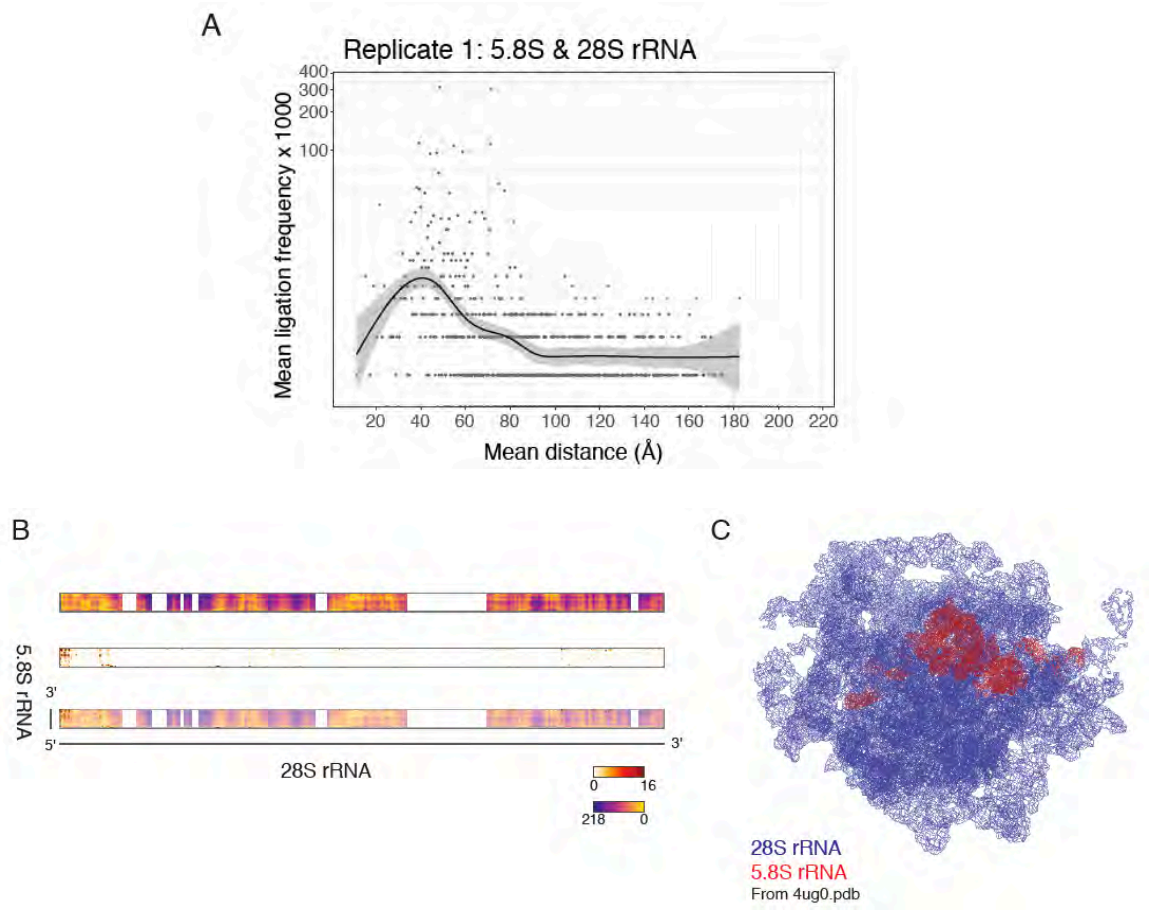


Figure 2.7. RIPPLiT captures expected inter-rRNA (5.8S-28S rRNAs) interactions

(A) Scatter plot showing mean ligation frequency as a function of Euclidean distance for 5.8S and 28S rRNA (replicate 1 + ligase) grouped into 1,000 equally-sized bins by number of chimeric junctions. Black line shows smoothing (generalized additive model (GAM)) with grey area displaying confidence interval (0.95) around smoothing.

(B) Heatmap of 3D distance between 28S and 5.8S rRNA (top) and Chimeric junction heatmap (middle) overlaid (bottom) for replicate 1.

(C) Structure of 28S rRNA (blue) and 5.8S rRNA (red) extracted from 4ug0.pdb.

With regard to Pol II transcripts, more than 80% of the >10,000 spliced species represented in our datasets had more than one unique intramolecular chimeric junction in the + ligase libraries (**Figure 2.9A, Table 2.3**). Further, because the number of unique intramolecular chimeric junctions per transcript was substantially higher in + than - ligase libraries (purple and green distributions), the vast majority of + ligase intramolecular chimeric junctions on Pol II transcripts represent bona fide ligation events. In contrast, intermolecular junctions were rare (only 5% of transcript pairs exhibiting apparent intermolecular junctions had more than one unique junction), and the cumulative histograms were nearly identical between the + and - ligase libraries (blue and pink distributions). This suggests that almost all intermolecular junctions obtained for Pol II transcripts in our datasets were due to mapping artifacts. Thus, in our clarified lysates from HEK293 cells we found no compelling evidence for biochemically stable intermolecular interactions between different mRNA species (**Figure 2.4D,E**; **Figure 2.8**).

Table 2.3. + ligase libraries have higher junction diversity (unique intramolecular chimeric junctions) than - ligase for both rRNAs and Pol II transcripts

# of unique junctions	Intra RNA	Libraries					
		Rep1 plus	Rep1 minus	Rep2 plus	Rep2 minus	Rep3 plus	Rep3 minus
rRNA	18S	36,354	403	26,976	353	17,213	314
	28S	19,569	694	11,433	528	5,897	356
	5.8S	232	37	121	9	151	13
	5S	3	0	3	0	1	0
Transcriptome	XIST	3,388	85	1,664	67	1,177	62
	UBR4	1,567	5	873	8	461	13
	TRRAP	1,002	6	524	10	268	12
	SPEN	929	13	497	11	251	7
	PRPF8	696	1	403	6	193	14
	EEF2	316	3	198	4	96	6

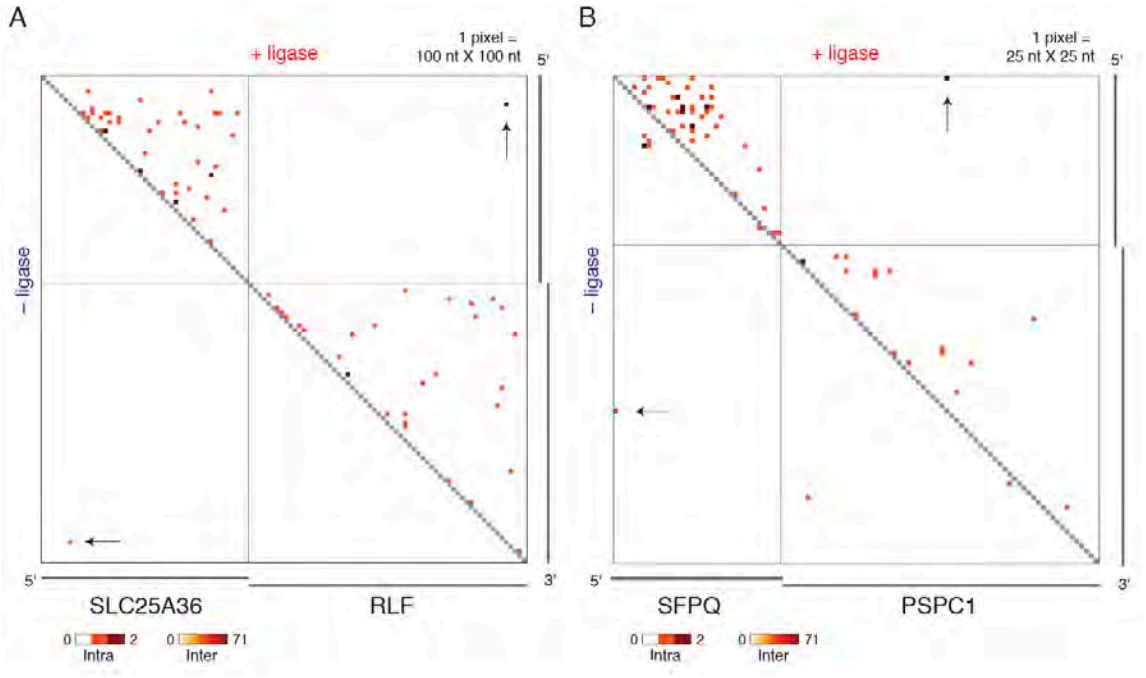


Figure 2.8. Intermolecular junctions captured in RIPPLiT are non-specific and ligase-independent, suggestive of mapping errors

(A,B) Chimeric junction heatmaps of representative intermolecular chimeric junctions. The transcript pairs shown (SLC25A36:RLF and SRPQ:PSPC1) were chosen for their high inter-RNA chimeric junction counts. Vertical and horizontal gray lines indicate transcript boundaries. Top left and bottom right heatmap quadrants contain intramolecular junctions; top right and bottom left quadrants contain intermolecular junctions. Arrows indicate inter-RNA chimeric junctions in both - and + ligase libraries. Note different color scales for inter- and intra-molecular chimeric junction frequencies.

RIPPLiT captures XIST structure

The most highly represented Pol II transcript in our datasets was XIST, a long non-coding RNA that functions in X chromosome inactivation. Consistent with its high abundance and nuclear retention (meaning that no EJCs are stripped away by cytoplasmic translation), XIST had ~2.5-fold more chimeric junctions than any other Pol II transcript. XIST consists of two unusually long terminal exons (~7 kb and ~11 kb) flanking four smaller (64 to 209 nt each) internal exons. Reflecting this primary structure, mapped fragments in both - and + ligase libraries exhibited the greatest depth on and around the four internal exons (**Figures 2.9B**). Chimeric junctions were ~7.5-fold more abundant in + than - ligase libraries, with + ligase junctions indicating the existence of short- (spanning <200 nt; typically occurring within one exon or between immediately adjacent exons), mid- (spanning 200-500 nt) and long- (spanning >500 nt) range interactions. Prominent long-range interactions occurred between the internal exons and the region immediately adjacent to the polyadenylation site in exon 6, and between the internal exons and a region close the 5′ end of exon 1 (**Figure 2.9B**; black arrows). This strong locus-specific pattern was highly reproducible across all three biological replicates (**Figure 2.9C**). PARIS, which uses in-cell psoralen crosslinking to identify base paired regions (Lu et al., 2016), has indicated the presence of two large secondary hairpin structures in XIST exons 1 and 6 (**Figure 2.9D**). Consistent with the existence of these structures, RIPPLiT captured interactions at the bases of these hairpins between regions otherwise distant in nucleotide space.

Another RNA secondary structure probing method, SHAPE-MaP (Smola et al., 2016), measures 2-OH accessibility and flexibility in cells. Compared to purified, protein-free XIST RNA folded in vitro, the region around the internal exons is substantially less reactive to the SHAPE reagent in cells (Lu et al., 2016; Smola et al., 2016). Because little secondary structure exists in this region, decreased SHAPE reactivity in cells is best explained by strong intracellular RNA-protein interactions, such as EJCs and associated factors. In this same region, RIPPLiT yielded numerous short- to medium-range chimeric junctions, suggestive of substantial RNA compaction. Combined, these datasets illuminate the higher order structure of XIST RNA in cells. Whereas PARIS reveals secondary structural elements and SHAPE-MaP reveals potential protein interaction domains, RIPPLiT reveals interactions between these features, including interactions between the internal exons and sequences flanking the hairpins. RIPPLiT thus provides information that both complements and extends the 3D structural information provided by other transcriptome-wide RNP structure probing methods.

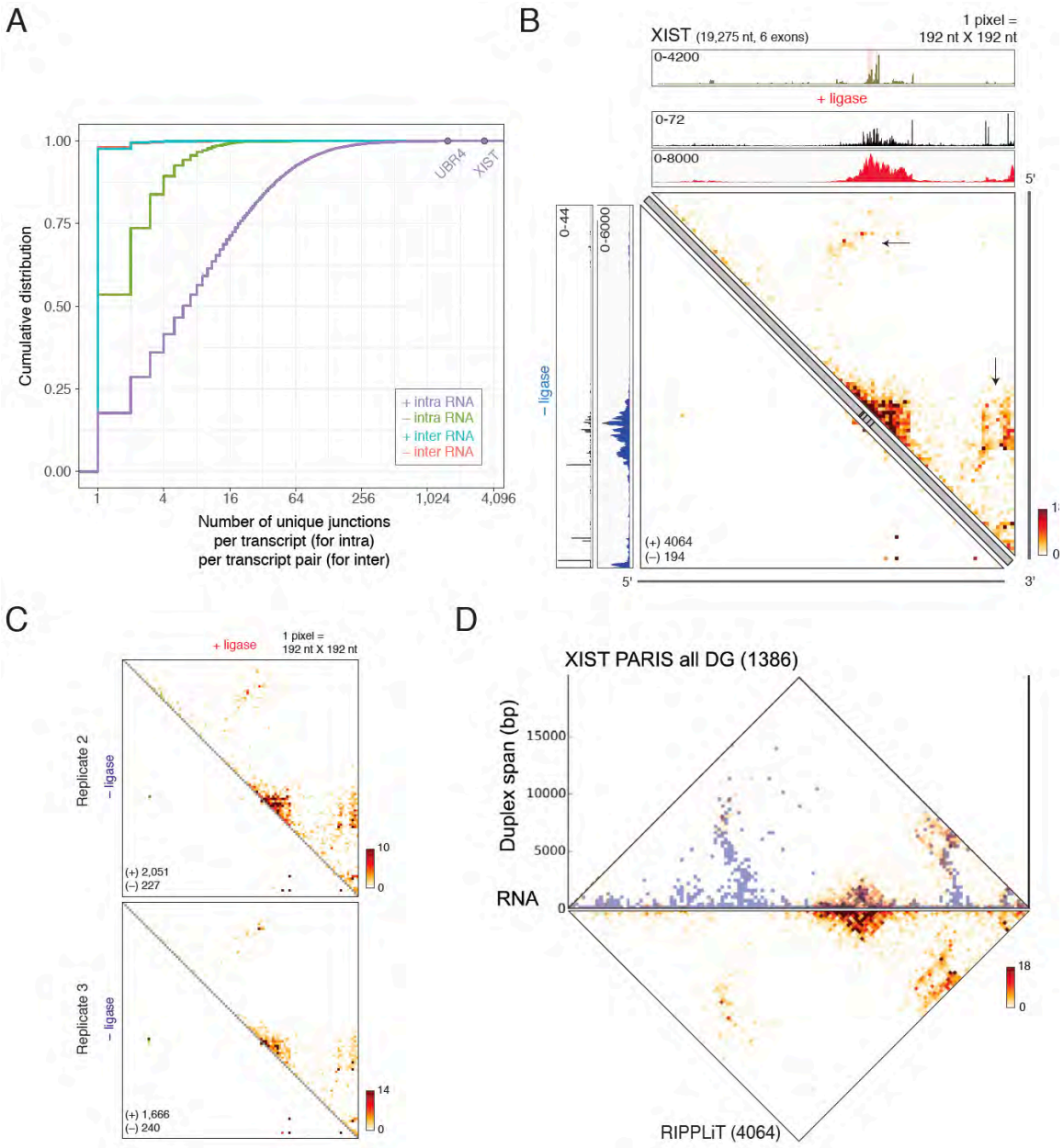


Figure 2.9. RIPPLiT captures higher-order structure of XIST

(A) Cumulative distribution of unique intramolecular and intermolecular chimeric junctions per RNA Pol II transcript. XIST and UBR4 marked for reference.

(B) Chimeric junction heatmaps for XIST (replicate 1). Red, blue and black coverage tracks as in **Figure 2.5**. Topmost coverage track (tan) displays short

EJC footprint coverage obtained from EJC RIPiT experiments (Singh et al., 2012), with pink lines indicating exon-exon junctions. Tick marks along diagonal gray bar, RefSeq-annotated exon-exon junctions in XIST. Prominent isolated chimeric junctions within the last exon in both - and + ligase heatmaps are due to unannotated minor alternative splice forms. Black arrows, long range interactions within the first and the last exon.

(C) XIST chimeric junction heatmaps for replicates 2 and 3.

(D) PARIS (Lu et al., 2016) Duplex Groups (DGs) overlaid with XIST RIPPLiT. Numbers in parentheses, number of junctions obtained in each dataset.

---

mRNAs lack strong locus-specific structures

We next focused on the set of 456 mRNAs exhibiting ≥100 intramolecular chimeric junctions in at least one biological replicate (**Figure 2.3E**, **Table 2.4**). In contrast to XIST, individual fragments in both + and - ligase libraries generally distributed across the entire length of spliced mRNAs. Exemplifying this were UBR4 (~16 kb; 106 exons), TRRAP (~12 kb; 70 exons), SPEN (~12 kb; 15 exons); PRPF8 (~7 kb; 43 exons), and EEF2 (~3 kb; 15 exons) (**Figure 2.10; Figure 2.11A**). On some mRNAs we observed a general enrichment of chimeric junctions toward the 5' end. This enrichment reflected an inherent 5' to 3' bias in fragment coverage (**Figure 2.11B**), which correlated equally with mRNA and gene length. The 5' end enrichment in chimeric junctions could be fully explained by the 5 end enrichment in fragment coverage, indicating that it was not due to any inherent 5 to 3 structural differences (**Figure 2.11C,D**). Mapping RIPPLiT fragments to the genome, however, revealed them to almost exclusively represent spliced exons (**Figure 2.11E**). Thus, our lysates contained a combination of nascent and fully mature mRNPs, with the nascent mRNPs likely released from chromatin as a result of sonication. By comparing fragment coverage near 5' and 3' ends, we estimate that for the longest mRNAs, the ratio of nascent to fully mature mRNPs is 3:2.

In rRNAs and XIST, strong chimeric junction enrichment between specific regions was indicative of locus-specific structures (arrows in **Figure 2.5A and Figure 2.9B**). In contrast, chimeric junctions on mRNAs distributed across the entire heatmap, with only a general inverse relationship between chimeric junction

density and nucleotide distance. This dispersed pattern, indicating a general absence of locus-specific structure, was observed in all three biological replicates (**Figure 2.12A**). This lack of locus-specific chimeric junctions was not simply attributable to read density differences, as XIST + ligase replicate 3 and UBR4 + ligase replicate 1 have a similar number of chimeric junctions, but very different heatmaps (**Figure 2.12B**). Notably, we did not observe elevated chimeric junctions spanning entire lengths of mRNAs, which would be expected if the 5' and 3' ends were closely juxtaposed in a circular arrangement.

Another notable feature was the coverage across an exceptionally large internal exon in SPEN (**Figure 2.10C**). Whereas the other internal SPEN exons ranged from 101 to 483 nt, exon 11 spans 8,176 nt. Nonetheless, mapped fragments and chimeric junctions were observed throughout this entire exon, with read density on this exon being ~12-fold greater than read densities on XIST exon 1, a segment of similar length. Thus, the structural scaffold of which the EJC is a part appears to encompass all internal mRNA exons regardless of length. Comparable decays in RIPPLiT and RNA-seq coverage on first and last exons indicate that this scaffold also extends into the terminal exons (**Figure 2.11F**).

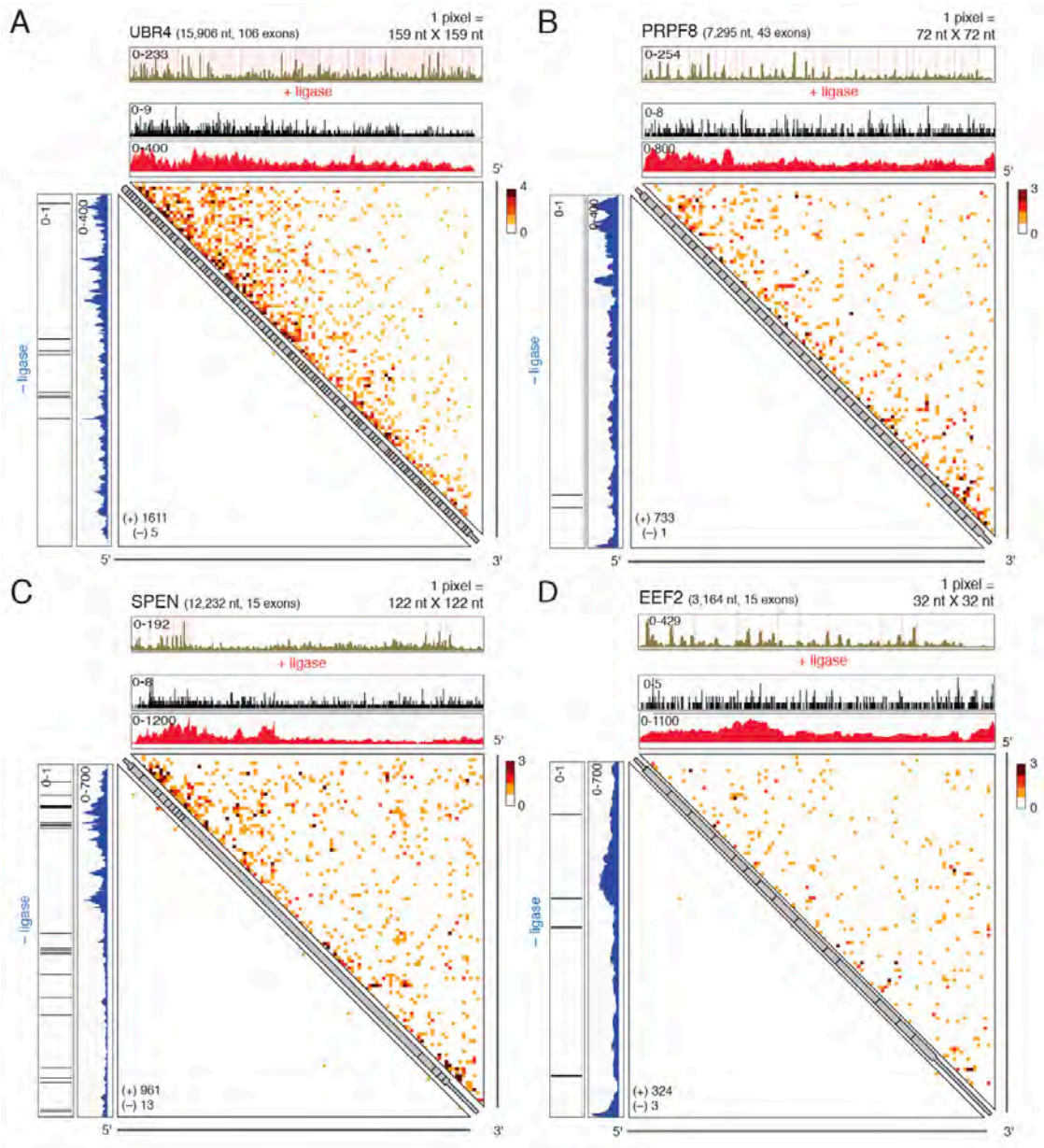


Figure 2.10. RIPPLiT captures higher-order structure of spliced Pol II transcripts

(A-D) Chimeric junction heatmaps for spliced mRNPs. All as in **Figure 2.9B** except that thicker and thinner sections of diagonal gray bar a coding exons and UTRs, respectively.

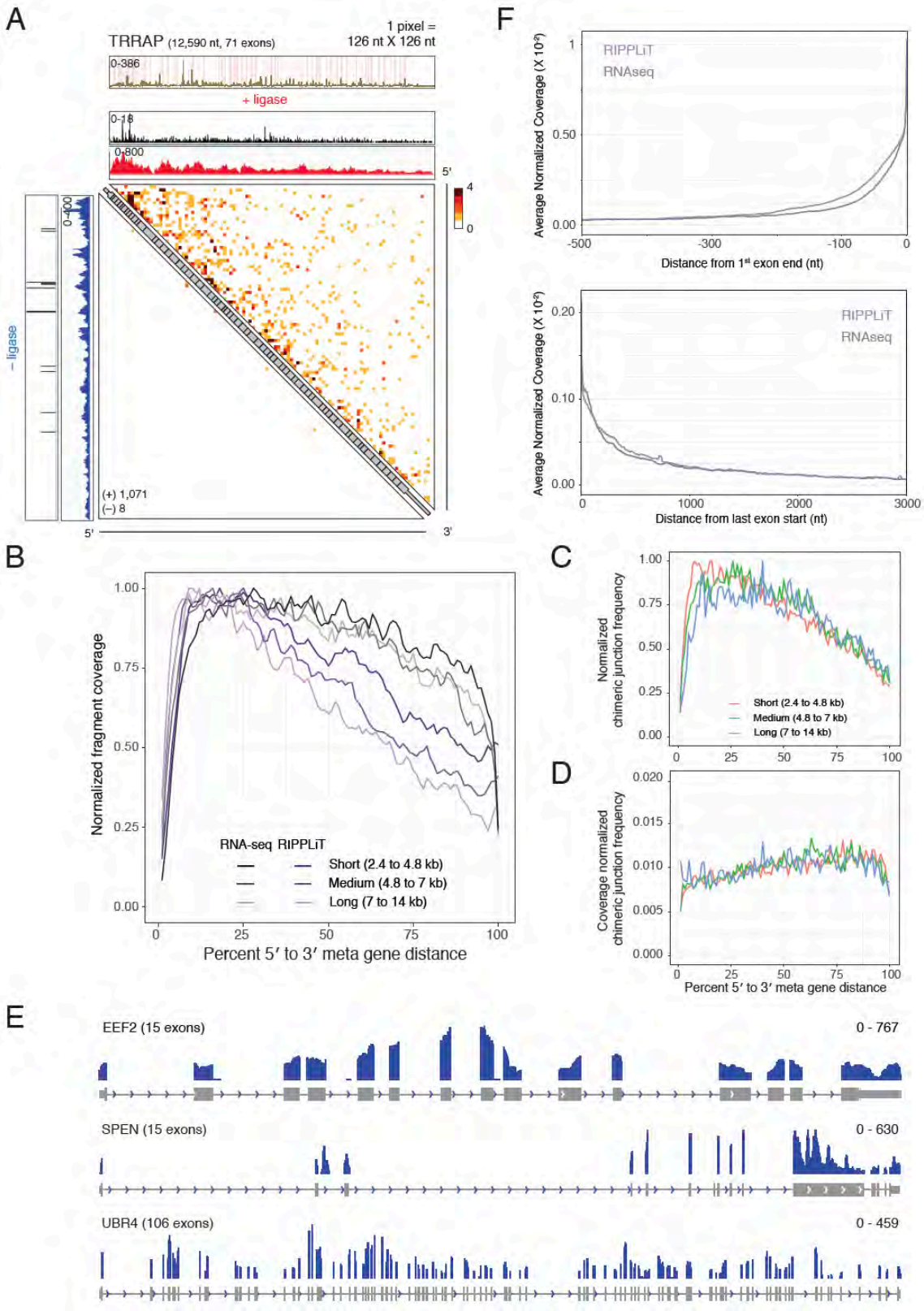

A
TRRAP (12,590 nt, 71 exons)
1 pixel =
126 nt X 126 nt
0-386
+ ligase
0-18
0-800
5'
4
0
0-400
– ligase
(+) 1,071
(–) 8
5'
3'
F
RIPPLiT
RNAseq
Average Normalized Coverage (X 10⁻³)
1
0.50
0.00
-500
-300
-100
0
Distance from 1st exon end (nt)
Average Normalized Coverage (X 10⁻³)
0.20
0.10
0.00
0
1000
2000
3000
Distance from last exon start (nt)
B
Normalized fragment coverage
1.00
0.75
0.50
0.25
0.00
RNA-seq RIPPLiT
Short (2.4 to 4.8 kb)
Medium (4.8 to 7 kb)
Long (7 to 14 kb)
0
25
50
75
100
Percent 5′ to 3′ meta gene distance
C
Normalized chimeric junction frequency
1.00
0.75
0.50
0.25
0.00
Short (2.4 to 4.8 kb)
Medium (4.8 to 7 kb)
Long (7 to 14 kb)
0
25
50
75
100
D
Coverage normalized chimeric junction frequency
0.020
0.015
0.010
0.005
0.000
0
25
50
75
100
Percent 5′ to 3′ meta gene distance
E
EEF2 (15 exons)
0 - 767
SPEN (15 exons)
0 - 630
UBR4 (106 exons)
0 - 459

Figure 2.11. RIPPLiT captures higher-order structure of spliced Pol II transcripts irrespective of their lengths

(A) Same as panels in **Figure 2.10**, but for TRRAP.

(B) Metagene plot showing normalized fragment coverage in RNA-seq (gray shades) and RIPPLiT (purple shades; replicate 1 + ligase) for the 456 mRNAs with ≥100 chimeric junctions in at least one biological replicate (**Figure 2.3C**) divided into three length groups as in **Figure 2.13**. Each transcript was divided into 100 bins of equal length and fragment coverage depth at each nucleotide within the bin summed. The fractional coverage for each bin was then calculated by dividing its sum by the sum of all bins. These fractional coverages were then summed across all transcripts in the indicated group and the resulting line plots individually normalized to the bin with the maximum value.

(C) Same as (B), except for chimeric junction frequency. Note different color scheme for the three length groups.

(D) Megagene plot for chimeric junction frequencies normalized to RIPPLiT fragment coverage. For this plot, all bins sum to 1 for each length group.

(E) Genome browser (IGV) screenshots showing replicate 1 - ligase fragment coverage on EEF2, SPEN and UBR4. RIPPLiT reads were mapped to GRCh37 using RSEM.

(F) Aggregate normalized coverage plot for RIPPLiT Replicate 1 (purple) and RNA-seq (Ge et al., 2016) (gray) libraries across first and last exon for the 456 mRNAs with ≥100 chimeric junctions in at least one biological replicate (**Figure**

2.3C). Data were normalized such that each transcript contributed equally to this plot.

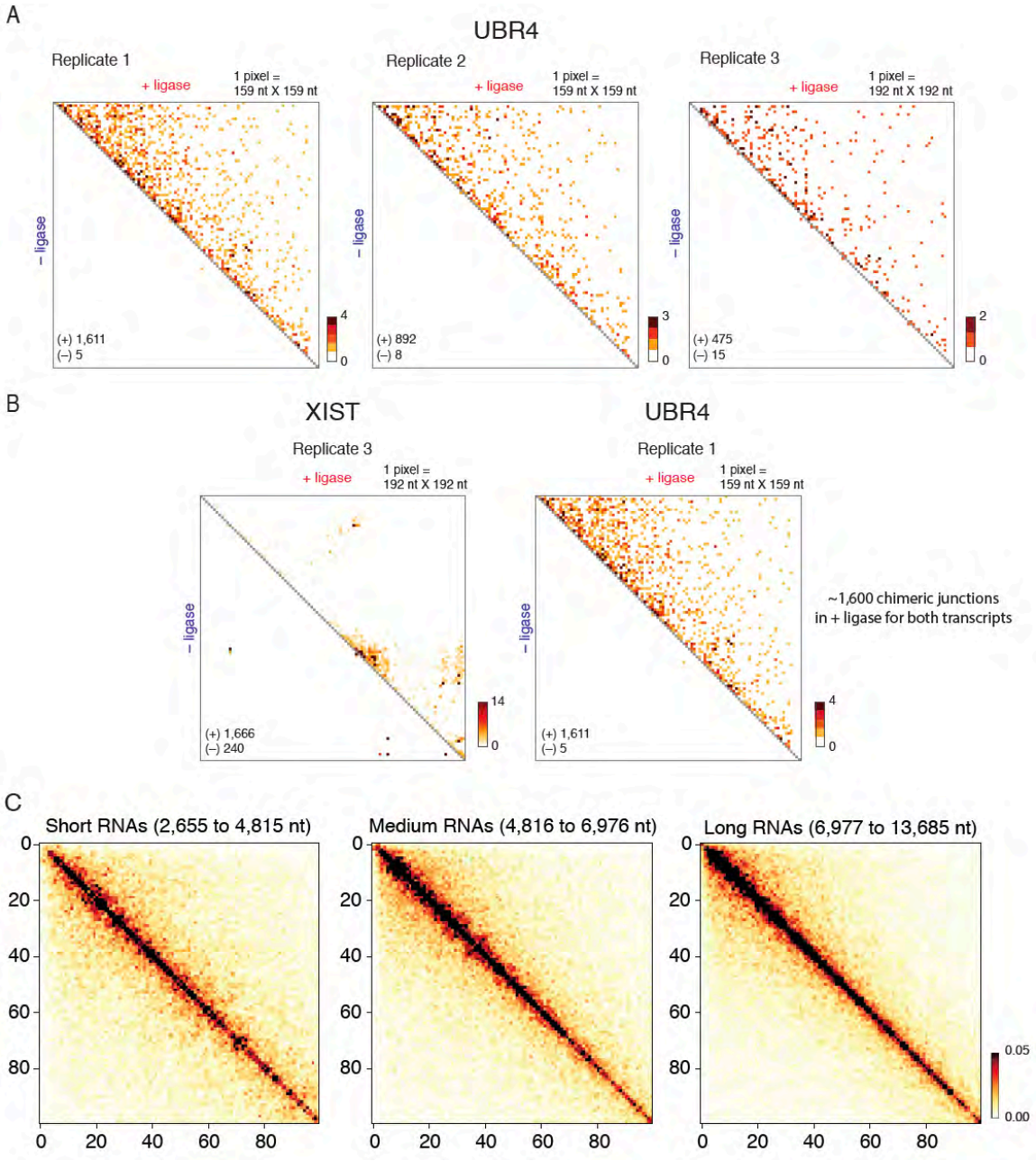


Figure 2.12. Heatmap pattern comparisons

(A) Chimeric junction heatmap for UBR4 across all biological replicates.

(B) Chimeric junction heatmap for XIST (replicate 3) and UBR4 (replicate 1), wherein both have ~1600 junctions in the + ligase library. Numbers in heatmap (lower left-hand corner): chimeric junctions obtained per indicated library.

(C) Uncorrected metagene heatmaps without balancing for transcripts of indicated length ranges. These heatmaps were created by dividing each transcript into 100 bins (1% length per bin = 1 pixel), normalizing each heatmap to a total chimeric junction frequency of 1.0, thus enabling data integration across multiple mRNA species of differing abundance. Color scale: Normalized mean chimeric junction frequency per pixel.

Polymer analysis indicates that pre-translational mRNPs form flexible rods

The relatively even distribution of chimeric junctions across the heatmaps indicates a general absence of strong locus-specific secondary structures within nascent and newly mature mRNPs. Rather our RIPPLiT data suggest that mRNAs behave more like flexible polymers. This is consistent with in-cell SHAPE (Spitale et al., 2015) and DMS-seq (Rouskin et al., 2014) data, which reveals that mRNAs in cells are more flexible (i.e., less structured) than in vitro-folded protein-free mRNAs. Such higher flexibility could reflect either loose packing of the mRNA strand or tight packing without specific points of contact. Without specific contact points provided by base pairing interactions, each mRNA molecule will adopt a slightly different mRNP structure in which different nucleotides are brought into close proximity, thus resulting in an even distribution of chimeric junctions in the ensemble population.

Another consistent feature of RIPPLiT data on mRNAs is the inverse relationship between ligation frequency and primary sequence span (**Figure 2.10A-D; Figure 2.11A**). Both the dispersed distribution of chimeric junctions across the entire length of mRNAs and the general decay in junction frequency with increasing nucleotide span are reminiscent of proximity ligation maps for chromatin. As with Hi-C interaction heatmaps for DNA (Lieberman-Aiden et al., 2009), polymer analysis of the ensemble data can provide insight into RNA packaging. Quantitative analysis of the relationship between contact frequency and the number of monomer units separating individual contact points (i.e.,

chimeric junctions in the case of RIPPLiT) can reveal the conformational organization of the polymer (e.g., whether or not it is a random coil) (Fudenberg and Mirny, 2012). We therefore calculated average chimeric junction frequency as a function of nucleotide distance for all 456 mRNAs for which we had sufficient data. To mitigate the effect of false positive chimeric junctions (those present in - ligase libraries) on our chimeric junction frequency quantification, we set the chimeric junctions bins in + ligase sample to 0 if they had any signal in the corresponding bin of - ligase library. To enable data aggregation across different length mRNA species, we divided each transcript into 100 bins (1% per bin) and calculated the fraction of total chimeric junctions in each bin. We could then create composite heatmaps by summing the junction frequencies across all bins for multiple mRNA species (**Figure 2.12C**).

Consistent with individual transcript heatmaps, we observed a homogenous spread of chimeric junctions throughout the composite heatmaps with a general decay in junction frequency with increasing distance. Since mRNA length varies widely, it seemed possible that mRNPs could adopt different structures at different length scales. To test this, we removed the most extreme length outliers (the 23 shortest and 23 longest = 10% of total), divided the remaining 412 transcripts into three equal member length groups (short: 2,655-4,815 nt; medium: 4,816-6,976 nt; and long: 6,977-13,685 nt) such that there is no more than 2-fold difference in length within each group. To control for the 5' to 3' bias in chimeric junctions driven by the mRNA and gene length-dependent 5' to 3' bias in fragment coverage, we

balanced the matrices using Iterative Correction as for Hi-C data analysis (Imakaev et al., 2012; Rao et al., 2014). Balancing equalizes the total number of interactions for each bin (Lajoie et al., 2015). This mitigates any biological (e.g., the locations of specific EJC and EJC-associated protein binding sites in any one mRNA species; differential fractions of nascent and fully mature mRNPs) or technical (e.g., read mappability) biases. The resultant composite heatmaps represent corrected average chimeric junction frequencies across many transcripts (**Figure 2.13A**). Strong similarity between all three heatmaps suggests no major structural differences driven by length. Further, as we had also observed for individual mRNAs (**Figure 2.10**), there was no evidence of stable interactions between the 5′ and 3′ ends (as indicated by the absence of signal in the upper right and lower left corners of heatmaps in **Figure 2.13A**).

The exponent (slope) by which the frequency of chimeric junctions decays with nucleotide or fractional distance (span) on log-log scaling plots can reveal properties of the folded state as well as the overall shape of the RNA polymer (**Figure 2.13B,C**; (Fudenberg and Mirny, 2012)). For instance, if mRNAs fold as simple free random coils, as has been proposed based on in-cell SHAPE data (Rouskin et al., 2014; Spitale et al., 2015), one would expect an exponent of -3/2, independent of mRNA length. Conversely, if mRNAs fully equilibrate to form equilibrium globules, proximity ligation frequency should initially decay with an exponent of -3/2, but then reach a plateau where ligation frequencies become independent of nucleotide distance (Fudenberg and Mirny, 2012; Lieberman-Aiden

et al., 2009). Neither feature was observable in our data, either for individual mRNAs (**Figure 2.13D; 2.14A**) or for composites of the three length groups (Figure **2.13E; 2.14B**). Thus, the paths assumed by mRNAs in nascent mRNPs are neither simple random walks nor equilibrium globules.

Regarding overall shape, a polymer with spatially proximal ends will exhibit high ligation frequencies at both the shortest and longest distances (**Figure 2.13C**). Conversely, a relatively rigid rod will display an initial decay in ligation frequency that suddenly drops off when no interactions become possible. A flexible rod-like polymer will exhibit high ligation frequency at short distances similar to the rigid rod, but since its flexibility enables longer range interactions, ligation frequency will decay much more gradually than for the rigid rod.

Three main features of the composite mRNA scaling plots are: (1) an initial slow decay (exponent ~-0.9) in ligation frequency for spans up to 40 nucleotides; (2) an even slower decay (exponent ~-0.4) for spans of 40-250 nucleotides; and then (3) a steady increase in the exponent as spans increase beyond 250 nucleotides. These features were independent of the most abundant spliced isoform level, indicating that alternative splicing is of minimal consequence for internal mRNP structure or overall shape (**Figure 2.14C**). These three features are remarkably similar to the Hi-C interaction patterns observed for prophase chromosomes (Gibcus et al., 2018), albeit on very different length scales. Such chromosomes are predicted to form highly packed and elongated structures containing numerous short (60-80 kb) loops connected by condensins. The initial

slow decay phase in the Hi-C scaling plots reflects interactions within individual loops, the second phase reflects contacts between adjacent loops, and the third phase (where the decay increases with span) is indicative of a flexible rod-like structure wherein interloop interactions become increasingly rare at very long distances.

For mRNAs, all three length classes exhibited identical first and second phases, with the very small exponents indicating a highly packed RNA conformation. The third phase was equally indistinguishable up to the point where the drop off sharply increased as spans began to exceed mRNA length in each class. This increasingly steep decay at larger nucleotide distances is indicative of a flexible rod. We therefore conclude that, irrespective of mRNA length or alternative processing, spliced nascent and mature mammalian mRNAs are compacted linearly into flexible non-circular rod-like structures (**Figure 2.13F**).

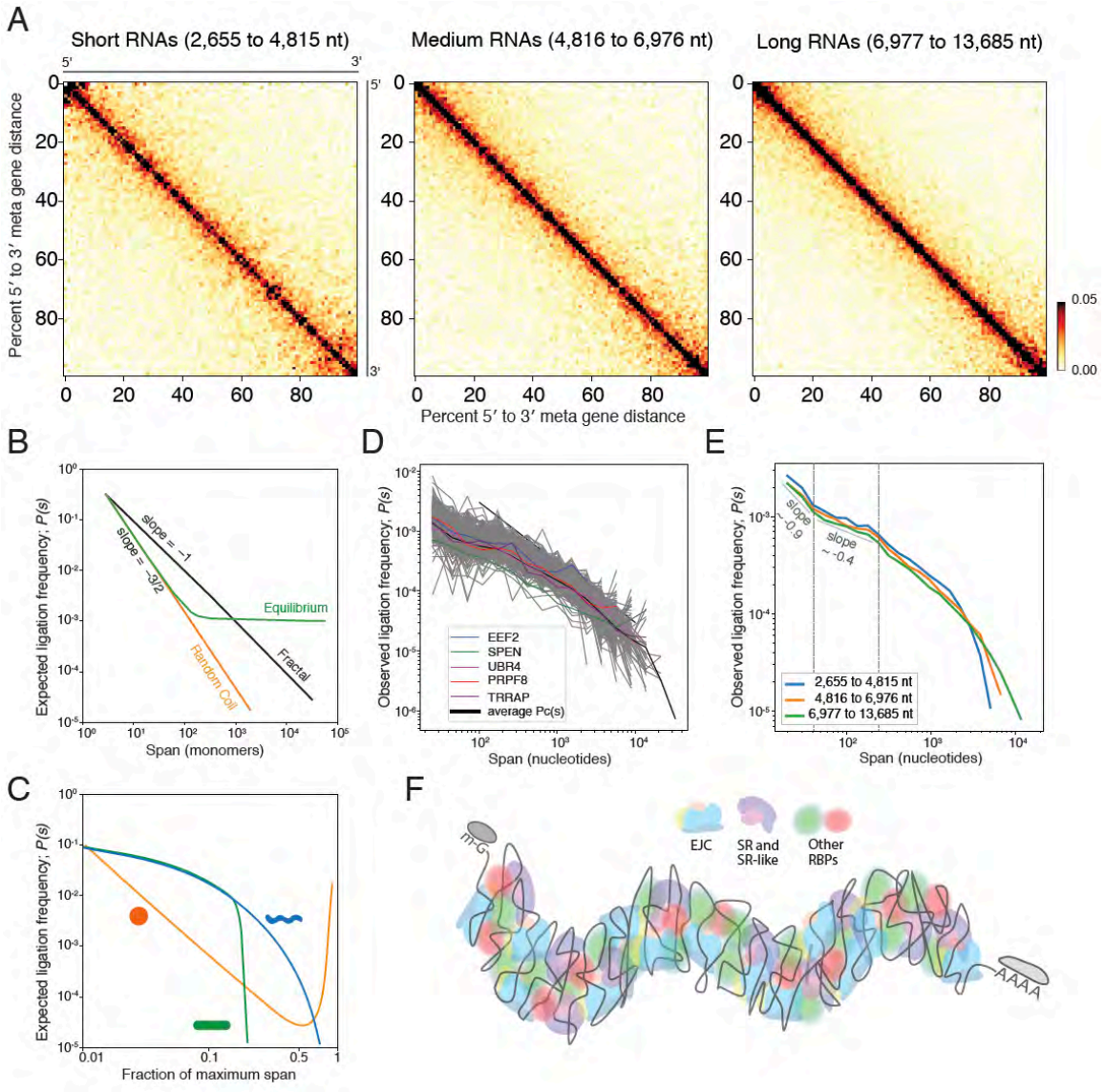


Figure 2.13. Within mRNPs, mRNAs are densely packed into linearly organized flexible rods

(A) Normalized metagene heatmaps (see text) for transcripts of indicated length ranges. Color scale: Normalized mean chimeric junction frequency per pixel.

(B) Expected scaling plots for different polymer types when contact probability, P(s) (equivalent to ligation frequency), is plotted against ligation span (distance in

number of monomers constituting the polymer) on log-log axes. EG, equilibrium globule; RC, random coil; FG, fractal globule (top).

(C) Expected scaling plots for different polymer shapes when P(s) is plotted against fraction of maximum span on log-log axes. Disc: globular polymer; Bar: rigid rod-like polymer; Worm: flexible rod-like polymer.

(D) Mean observed scaling plot for 456 transcripts (grey lines) with more than 100 chimeric junctions in at least one biological replicate. Data from all replicates were combined and mean ligation frequency plotted as a function of distance in nucleotides on log-log axes. Each color represents a transcript (exemplifying different lengths) shown in **Figure 2.10** and **Figure 2.11A**. Black line indicates a slope of -1.

(E) Mean observed scaling plot for each length group shown in **Figure 2.13A** with all replicates combined. Vertical dotted lines indicate approximate boundaries for exponent changes.

(F) Model of rod-like mRNP, with EJCs and other mRNP proteins protecting large RNA regions from nuclease digestion.

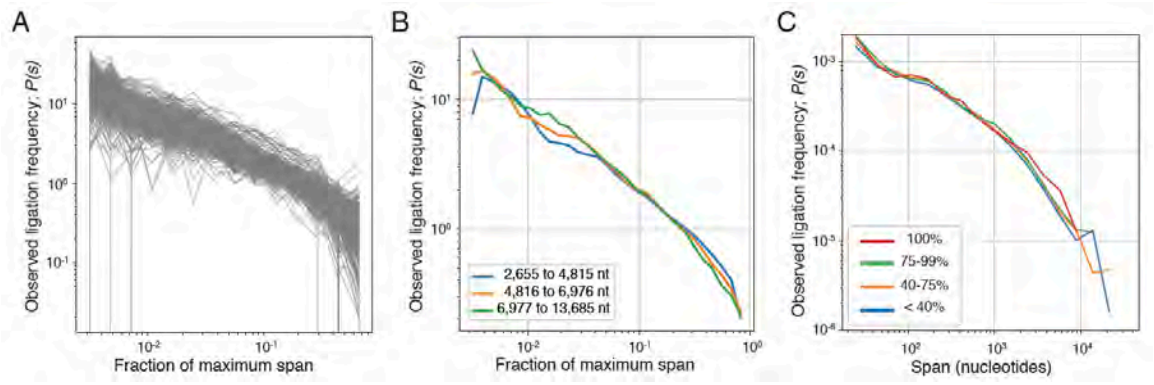


Figure 2.14. Transcript scaling plots are unaffected by transcript length or alternative isoform expression

(A) Same as in **Figure 2.13D** but with fraction of maximum span plotted on X-axis.

(B) Same as in **Figure 2.13E** but with fraction of maximum span plotted on X-axis.

(C) Mean observed scaling plot for transcripts grouped by the most abundant isoform expression level. Mapping and alternate isoform level estimation with RSEM revealed that 13 of our 456 genes (same as in **Figure 2.13**) have only one expressed mRNA isoform in our datasets (Group 1). The remaining 443 were divided into three groups (Group 2: 101 genes in which the most abundant isoform is between 75 and 99%; Group 3: 227 genes in which the most abundant isoform is between 40 and 75% of transcripts; and Group 4: 115 genes in which the most abundant isoform is < 40% of transcripts). Data from all 3 replicates were combined.

DISCUSSION

Here we describe a new proximity ligation technique, RIPPLiT, to capture higher order structural information in RNPs of defined protein composition. By enriching for EJC-associated RNAs, we were able to investigate the overall architecture of spliced Pol II transcripts in association with stably-bound proteins acquired during transcription and pre-mRNA processing. Our data indicate that prior to translation, spliced mRNAs and their associated proteins are tightly packed into linearly-organized flexible rod-like structures irrespective of length.

RIPPLiT captures 3D RNP structural information independent of base pairing

Proximity ligation coupled with semi-quantitative PCR has been used to map chromatin interactions and to delineate the 3D conformation of yeast chromosomes (Dekker et al., 2002). Subsequent technical developments allowing identification by proximity ligation of chromatin interactions genome-wide enable interrogation of the overall 3D arrangement of DNA in cells, and such studies are starting to reveal the folding principles of chromatin in different biological states at high resolution (Gibcus et al., 2018; Lieberman-Aiden et al., 2009). The first adaptation of proximity ligation to RNA was CLASH (Cross-linking, Ligation, And Sequencing of Hybrids; (Kudla et al., 2011)), which captured intermolecular RNA-RNA base pairing (e.g., miRNA-mRNA) associated with a particular UV-crosslinked RBP (e.g., Ago1; (Helwak et al., 2013)). Multiple subsequent studies have described methods for capturing intra- and intermolecular base-pairing

transcriptome-wide, either by psoralen crosslinking (e.g., PARIS (Lu et al., 2016), SPLASH (Aw et al., 2016) and LIGR-Seq (Sharma et al., 2016)) or by UV-crosslinking to the double-stranded RBP, Staufen 1 (HiCLIP; (Sugimoto et al., 2015)). Base-pairing interactions also dominate datasets obtained from other proximity ligation methods (RPL (Ramani et al., 2015) and MARIO (Nguyen et al., 2016)) even though these methods were not specifically designed to capture secondary structures. Therefore, previous proximity ligation methods for probing higher order structures within RNPs mainly captured interactions driven by base pairing, and so were highly skewed toward ncRNAs. In contrast, we enriched for EJC cores, deposition of which requires single-stranded RNA (Andersen et al., 2006; Mishler et al., 2008; Singh et al., 2012). Although our data do include some interactions driven by base-pairing (e.g., those at the base of large secondary structures in XIST; arrows in **Figure 2.9B**), more common were chimeric junctions in regions without strong base-pairing tendencies. Thus, a major difference between previous proximity ligation methods and RIPPLiT is that we capture higher order structure in RNPs formed as a consequence of RBP binding and packaging as well as base-pairing. This coupled with the high coverage throughout mRNAs (**Figure 2.10**) allows for interrogation of the overall 3D organization of stable RNA-protein complexes.

The strong enrichment of our EJC RIPPLiT datasets for spliced Pol II transcripts and the high depth of chimeric junctions in the + ligase libraries (464,426 unique junctions on Pol II transcripts in all) allowed us to investigate the

3D structural properties of hundreds of spliced mRNAs. Notably, the chimeric junctions we observed on mRNAs were almost entirely intramolecular (**Figure 2.9A**). Even though intermolecular junctions between rRNAs were readily apparent in our datasets (**Figure 2.7**), we could detect no specific multi-RNA mRNPs (**Figures 2.4D,E; Figure 2.8**). Nonetheless, intermolecular interactions between different mRNA species have been reported in previous proximity ligation studies that employed in situ crosslinking in yeast, HeLa cells, lymphoblastoid cells, hESC line H1 (Aw et al., 2016) and mouse embryonic stem (ES) cells (Nguyen et al., 2016), albeit at extremely low counts per ligation junction. Thus, if intermolecular mRNA complexes do occur in HEK293 cells, they are either exceedingly rare or they cannot survive the clarification, T1 RNase digestion and biochemical purification conditions employed here. We note that the 15,000 g centrifugation step we used to clarify our cell lysates likely eliminated any very large RNP granules (e.g., P bodies) (Hubstenberger et al., 2017). In the future, it would be of great interest to modify RIPPLiT conditions to allow for interrogation of larger RNP complexes, such as the multi-mRNA transport granules observable in cells with highly extended processes (e.g. neurons and oligodendrocytes) (Carson et al., 2008; Park et al., 2014).

Structural features of mRNPs revealed by RIPPLiT

Our previous study of EJC footprints on spliced Pol II transcripts focused primarily on short (12 to 25 nt) footprints expected for monomeric EJCs (Singh et

al., 2012). In that study, we subjected EJC-bound RNAs to high concentrations of RNase I. Yet despite this stringent digestion, the predominant fragment sizes obtained were 30-150 nt, much longer than the EJC or any other known RBP footprint. This indicated that, when associated with their complement of stably-bound proteins, spliced Pol II transcripts are highly protected from RNase digestion.

In the current study we subjected EJC-containing RNPs to milder RNase T1 digestion, both with the intent of preserving overall RNP structure and to generate single-stranded RNA ends with sufficient toeholds for T4 RNA ligase I. For spliced mRNAs, we observed fragment coverage over almost the entire transcript (**Figures 2.10; Figure 2.11**). This is consistent with previous mapping data indicating canonical EJC binding upstream of exon junctions plus strong association with other non-canonical sites, most likely through protein-protein interactions with other RNA-binding proteins (Sauliere et al., 2012; Singh et al., 2012). That the fragment coverage extends even across extremely long internal exons (**Figure 2.13F**) indicates that canonical EJCs are just one small part of a much larger, highly stable multicomponent interaction network.

Both individual (**Figure 2.10A,C; Figure 2.11A**) and uncorrected composite (**Figure 2.12C**) heatmaps revealed 5' to 3' biases in fragment coverage (**Figure 2.11B**) and chimeric junction frequency (**Figure 2.11C**), both of which were a function of mRNA length. Because a similar bias was observed for SPLASH, an IP-independent RNA-proximity ligation method (Aw et al., 2016), this effect is not

specific to RIPPLiT. In both procedures it most likely reflects capture of some nascent (i.e., incompletely transcribed) mRNPs, which in our protocol may have sheared from the chromatin as a consequence of sonication. Thus, our data contain a mixture of nascent and mature mRNPs. However, iterative correction of composite heatmaps, which controls for any differences in region-specific fragment coverage (Imakaev et al., 2012; Lajoie et al., 2015), revealed similar patterns of chimeric junction spans along the entire lengths of three different mRNA length classes (**Figure 2.13A**). This is consistent with the idea that mRNA packaging occurs co-transcriptionally and suggests that there are no major changes to the overall RNA polymer arrangement upon chromatin release.

Scaling plots (**Figure 2.13**) indicate that mammalian pre-translational mRNAs exist as linearly organized rod-like structures. Within these rods, EJCs and associated proteins likely form a stable scaffolding that nucleates and maintains the mRNA in a densely packaged state. Our current data combined with previous imaging studies and our findings that mRNPs are strongly protected from RNase digestion suggest a very compact structure with many different species contributing to a dense network of protein-RNA and protein-protein interactions. The high frequency of chimeric junctions with spans less than 40 nt likely represents ligations occurring between the ends of short excursions from this tight proteinaceous core.

Rod-like packaging, likely occurring in a first-come-first-served order during mRNA synthesis and processing, has multiple biophysical and mechanical

advantages. First, by preventing RNA knots and by compressing exons into a form less prone to physical breakage than extended RNA, compacted rods can preserve the functionality and integrity of newly synthesized messages. Second, rod-like nanoparticles diffuse more rapidly than spherical particles through adhesive polymeric gels, such as provided by the dense polymeric environment of the nucleus with its innumerable weak interaction sites for RNA-protein complexes (Wang et al., 2018). Third, rods of uniform thickness rather than spheres of varying diameter dependent on mRNA length have clear advantages for passage through the nuclear pore complex. Intriguingly, a recent intracellular tracking study following differently shaped nanoparticles reported that even when equal diameter objects are compared, rod- and worm-shaped nanoparticles traverse nuclear pores more efficiently than spheres (Hinde et al., 2017). Consistent with this, even globular Balbiani ring mRNPs can be seen passing through nuclear pore as rods (Skoglund et al., 1983).

Finally, mRNAs undergoing translation often have functional interactions between the 5' and 3' ends (i.e., between proteins bound to the cap and polyA tail) (Archer et al., 2015; Christensen et al., 1987; Wells et al., 1998). However, we observed no evidence for such interactions in pre-translational mRNPs. That the ends of mammalian mRNPs do not interact prior to translation is supported by a recent in cell single molecule FISH study, which found no colocalization in the nucleus of probes hybridizing to 5' and 3' ends of three different mRNAs (Adivarahan et al., 2017). Interestingly, in that study, 5'-3' end interactions were

observable in the cytoplasm only after polysome collapse upon puromycin treatment. Another study performed on specific mRNAs in yeast showed that rather than being a global phenomenon, multiple factors affect mRNA circularization including phase of translation and the mRNA sequence itself (Archer et al., 2015). Our data are consistent with the hypothesis that functional mRNA circularization occurs primarily within polysomes, not in pre-translational mRNPs.

EXPERIMENTAL METHODS

Experimental model and subject details

Human embryonic kidney HEK293 (female) cells stably expressing near endogenous levels of a FLAG-tagged Magoh protein were maintained in Dulbecco's modified eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin, and grown at 37 °C.

EJC RIPPLiT

For EJC RIPPLiT, TRex-HEK293 cells containing a stable copy of FLAG-tagged Magoh under control of a tetracycline-inducible promoter were grown in six 15 cm plates. FLAG-tagged Magoh expression was induced by treating cells with tetracycline for 16-20 hrs. To limit EJC removal by the pioneer round of translation,

harringtonine was added to 2 ng/mL one hour prior to harvest. Unless otherwise noted, all subsequent steps were performed on ice using ice-cold buffers. Cells were rinsed once with 10 mL phosphate-buffered saline (PBS) containing 2 ng/mL harringtonine, harvested by scraping, and lysed for 10 min in 8 mL HLB [Hypotonic Lysis Buffer: 20 mM Tris-HCl pH7.5, 15 mM NaCl, 10 mM EDTA, 0.5% NP-40, 0.1% Triton-X-100, protease inhibitor cocktail (Roche)]. Lysates were sonicated (Branson Digital Sonifier-250) in aliquots of 4 mL each, at 40% amplitude using a Microtip for a total of 16 s (in 2 s bursts with 10 s intervals). After adjusting NaCl to 300 mM, recombined lysates were clarified by centrifugation at 15,000 xg for 10 min at 4 °C and then diluted to 18 mL in HBL containing 300 mM NaCl. The diluted lysate was incubated for 2 hr at 4 °C with 750 µL anti-FLAG agarose beads (50% slurry, Sigma) pre-washed twice with 10 ml IsoWB [Isotonic Wash Buffer: 20 mM Tris-HCl pH7.5, 150 mM NaCl, 0.5% NP-40]. RNP complexes captured on beads were washed 4x with 10 mL IsoWB. After the fourth wash, beads were transferred to a new tube and incubated for 10 min at 37 °C with intermittent shaking (ThermoMixer) with one bed volume IsoWB containing 10 Units/mL RNase T1 (Life Technologies, EN0541). After washing beads 4x with 10 mL IsoWB to remove RNase T1, FLAG-epitope containing complexes were affinity eluted into one bed volume of IsoWB containing 250 µg/mL FLAG peptide by gentle shaking at 4 °C for 2 hr. The recovered eluate was used as input for a second IP by increasing its volume to 1 mL with ILB [Isotonic Lysis Buffer: HLB containing 150 mM NaCl]. This suspension was incubated with Bethyl A302-980A anti-eIF4AIII antibody (3 µL per

15 cm plate) that had been pre-coupled to ProteinG-Dyna-beads (35 µL per 15 cm plate) according to manufacturer's (Invitrogen) instructions. Immunoprecipitation was carried out at 4 °C for 2 hr. Beads with captured RNP complexes were washed 6X with 1 mL IsoWB, with a tube change between the 3rd and 4th washes.

RNase T1 digestion leaves 3′ phosphate and 5′ hydroxyl ends which need to be converted to 3′ hydroxyl and 5′ phosphate ends for ligation reactions. To accomplish this, beads were washed 3x with polynucleotide kinase (PNK) wash buffer [70 mM Tris-HCl, 10 mM $MgCl_2$, 5 mM DTT, 0.5% NP40] and evenly distributed between two 2 ml Eppendorf tubes. Following addition of Dephosphorylation Reaction Mix [1X PNK buffer, 2.5 U T4 PNK (NEB, M0201), 1 mM DTT] to a final volume of 50 µL and a 45 min incubation at 37 °C, Phosphorylation Mix [1X PNK buffer (70 mM Tris-HCl, 10 mM MgCl2, no DTT), 1 mM ATP, 1 U T4 PNK] was added to the same tubes and reactions additionally incubated at 37C for 1hr. For the 5′ end protection assay (**Figure 2.2**), Hot Phosphorylation Mix [1X PNK buffer, 150 µCi g-^{32}P-ATP (Perkin Elmer, NEG035C005MC), 1 U T4 PNK] was added instead; after 20 min at 37 °C, the reaction was supplemented with 1 mM cold ATP and further incubated for another 20 min at 37 °C. In all cases, beads harboring end-repaired RNP complexes were combined into a single tube and washed 3x with PNK wash buffer before proceeding to the ligation step.

For ligation reactions, beads were evenly distributed among four 2 ml Eppendorf tubes, two for + ligase and two for the - ligase control. To each + ligase

tube, Ligation Mix [1X T4 RNA ligase buffer, 1 mM ATP, 50 U T4 RNA ligase 1 (NEB, M0204)] was added a final volume of 50 µL; - ligase reactions were identical, except no ligase was added. All additions of mixtures were done on ice and then reactions incubated at 25 °C overnight. After consolidating the reactions into a single tube for each condition, beads were washed 3x with ice-cold IsoWB and eluted with 40 µL of Clear Sample Buffer [100 mM Tris-HCl pH 6.8, 4% SDS, 10 mM EDTA, 100 mM DTT], first at 25 °C for 5 min, and then at 95 °C for 2 min.

RNAs were recovered from the eluate by adjusting the volume to 300 µL with water, extracting twice with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1; pH 4.5) and once with chloroform:isoamyl alcohol (24:1). The recovered aqueous phase was supplemented with 10 µg of glycogen, 300 mM sodium acetate (pH 5.2) and RNA was precipitated with 3 volumes of 100% ethanol overnight at -20 °C. After centrifugation, the pellet was washed with 70% ethanol and dried. RNA was re-suspended in 20 µL of water and stored at -80 °C for subsequent steps. 1/20th of each sample was run on a Bioanalyzer to visualize the RNA profile (**Figure 2.1B**).

5′-end protection assay

RNAs that were ^{32}P-labelled during the 5′ phosphorylation reactions were extracted from + and - ligase samples and a portion of each was incubated at 37°C for 1 hr with Calf Intestinal Phosphatase (CIP, M0290) Mix [1X NEB buffer 2, 20 U

CIP]. Another portion was treated identically, except that in the absence of CIP. Following ethanol precipitation in the presence of glycogen, RNAs were resolved on a 26% UREA-PAGE gel and radioactive species detected by phosphorimaging (**Figure 2.2**).

Library preparation and sequencing

Deep sequencing libraries were prepared from three biological replicates of RIPPLiT experiments performed on three different days using the SMARTer smRNA-Seq Kit (Clontech, 635030) as directed and starting with 200 ng RNA per sample. After 12-14 amplification cycles, PCR products containing 200 to 400 nt inserts were size-selected using a Pippin HT. Each library contained a unique barcode, enabling all six libraries (+ and - ligase for 3 biological replicates) to be mixed together and sequenced in a single run on an Illumina NextSeq instrument using the NextSeq 500 version 2 (Illumina, Inc., FC-404-2004) paired-end sequencing kit. To maximize read depth from the + ligase samples, + and - ligase libraries were mixed at a 2:1 ratio.

Read pre-processing and sequence alignment

More than 80% of RIPPLiT read pairs overlapped by ≥10 nt and so could be merged into a single read using Paired-End AssembleR (pear, v0.9.6) (Zhang

et al., 2014) with its default options. Paired-end reads with insufficient overlap were discarded. Reads were further processed by removing any parts of the sequencing adaptor using Cutadapt (v. 1.7.1) (Martin, 2011). The SMARTer-seq protocol adds 3 untemplated nucleotides at the 5′ end and a poly-A at the 3′ end which also need to be trimmed. Therefore, 3 nucleotides at the 5′-end and 15 nucleotides at the 3′ end were trimmed using Cutadapt (v. 1.7.1)

Because many of the + ligase chimeric reads contained more than one fragment junction and inverted fragment junctions occurred with near equal frequency as inline junctions, no existing alignment algorithm capable of mapping these chimeric products. Therefore, we developed ChimeraTie, a bioinformatics tool that uses the local alignment mode of Bowtie2 (alignment parameters: --local -D 20 -R 3 -N 0 -L 20 -i S,1,0.50 -p10 --rdg 5,6 --rfg 5,6) to iteratively map all fragments within a single read (**Figure 2.1D**). Because of the high abundance of rRNAs and snRNAs in our libraries, we iteratively mapped reads first to rRNAs, then to snRNAs and then to a locally-generated HEK293 transcriptome. This transcriptome was created by mapping non-rRNA/snRNA mapping reads from a -ligase library to the hg37 genome using RSEM (Bowtie2 and default parameters) (Li and Dewey, 2011); the highest abundance isoform was chosen as the single reference transcript for each gene.

To increase mapping stringency, additional filtering steps were implemented post-alignment. These included eliminating all non-uniquely mapping fragments (XS tag), those mapping antisense to the reference transcriptome (SAM

flag 16), and any fragment alignment containing >3 mismatches (XM tag) or a gap of >3 nt (XG tag). The surviving reads containing multiple fragments were converted into an interaction file containing pairwise arrangement information for all fragments within the read. As a final filtering step, we selected only pairs of fragments that directly abut one another in the read ("Direct" junctions). This file enabled us to calculate fragment junction frequencies per transcript (intra-RNA fragment junctions) or between transcript pairs (inter-RNA fragment junctions). It also served as the basis for creating fragment junction matrices for individual transcripts that can be converted into heatmaps using the heatmap.pl script from the cWorld program (https://github.com/dekkerlab/cworld-dekker).

Methods for P(s) and heatmaps

To obtain average contact matrices for mRNAs, we started with contact maps binned at 10 nt resolution. To eliminate false positive chimeric junctions (i.e., not due to ligation), each matrix from + ligase library was filtered to set bins to 0, if the corresponding bin in the - ligase library had a non-zero junction frequency. We also filtered the first diagonal of the matrix to remove any chimeric junction spanning 10 or fewer nucleotides. For combined replicate analyses, we added together contact maps from all replicates for each mRNA. We rescaled each matrix to a 100x100 matrix using Python function mirnyliib.numutils.zoomArray, which is a wrapper around scipy.ndimage.zoom that allows for better down-sizing. To

ensure that each mRNA contributes equally to the average, we divided each rescaled map by its mean. The maps were then added together to yield an average contact map. Each average map was then balanced using Iterative Correction and normalized such that each row sums to 1.

To obtain average P(s) curves, we started with contact maps binned at 20 nt resolution. For each map, we masked out rows and columns with zero recorded interactions. We then calculated the sum of each diagonal O(s), and the number of non-masked elements in each diagonal E(s). To ensure smooth P(s) curves at large spans, O(s) and E(s) were aggregated in log-spaced integer bins with the difference of 1.1, obtained using mirnylib.numutils.logbins (1,4000,1.1). We then calculated P(s) for each mRNA by dividing aggregated $O^{agg}(s)/E^{agg}(s)$. To ensure that each mRNA has equal contribution to the average P(s) curve, each P(s) curve was normalized to sum to one. Average P(s) curves for the indicated length groups were then calculated from these normalized individual mRNA P(s) curves.

Data availability

The accession number for the RIPPLiT raw fastq files and processed interaction files is: GSE115788

Endnotes for Chapter II

Chapter II is adapted from the following:

Metkar M, Ozadam H, Lajoie B, Imakaev M, Mirny L, Dekker J, Moore M. J. (2018). Higher-order organization principles of pre-translational mRNPs. (Pre-accepted at Molecular Cell).

Author Contributions

M.M. and M.J.M. originally conceived the project, with M.M. executing all wet bench experiments. B.R.L. and H.O. conceived and wrote ChimeraTie and M.M. implemented all computational analyses with input from M.J.M and J.D. Polymer analysis was done by M.I. under the guidance of J.D. and L.A.M. All authors contributed to data analysis and interpretation. With input from all authors, M.M., J.D. and M.J.M. were primarily responsible for writing the paper.

Chapter III

RIPPLiT and ChimeraTie: Methods to capture higher order structures of mRNPs

Abstract

Development of high-throughput approaches to map RNA interactomes is providing new insights into gene expression regulation through identifying novel RNA-RNA interactions as well as better understanding of fundamental principles of RNP life-cycle such as mRNP shape inside cells. To capture transcriptome wide RNA interactomes associated with ribonucleoprotein (RNP) complexes of defined protein content, we have developed a novel method called RIPPLiT (RNA ImmunoPrecipitation and Proximity Ligation in Tandem). Further, we developed a new bioinformatics tool, ChimeraTie, to map, analyze and visualize the complex chimeric reads obtained as a result of RIPPLiT. RIPPLiT can potentially be applied to understand base-pairing mediated interactions as well as protein mediated RNA-folding interactions for a particular RBP or an RBP complex. While, ChimeraTie can be applied to any biochemical approach that results in chimeric reads such as those obtained by proximity ligations, alternative splicing or RNA circularization. Thus, allowing the study of RNP complexes throughout their life cycle.

Introduction

The past few decades have shown that RNAs have multiple levels of encoded information apart from the primary sequence. These are exemplified by secondary structural features like base-pairing (e.g., iron-responsive element (Piccinelli and Samuelsson, 2007)), specific motifs (e.g., AU-rich elements (AREs); (Brennan and Steitz, 2001)), or specific (e.g., HuR; (Brennan and Steitz, 2001))/ non-specific (e.g., Exon-junction complex (EJC); (Le Hir et al., 2000)) RNA-binding proteins sites or even inter-RNA interactions (e.g., miRNA-mRNA interactions) (Helwak et al., 2013; Lee et al., 1993). During all of the pre-translational processes, RNAs need to be packaged in a way that they are accessible for decoding this information through various factors but also protected from degradation and thus maintain their fidelity. However, this regulation of messages through inherent messenger 3-dimensional organization has largely been overlooked due to the lack of robust techniques that allow the capture of this information as well as the dynamicity of the RNP complexes.

Although only a few studies have looked at 3D structure of RNAs, a large focus has been put in understanding RNA secondary structure. To this end various methods were developed to capture the *in vivo* and *in vitro* secondary structure. The first generation of techniques (reviewed in (Piao et al., 2017)) developed used nucleases that either cleave single stranded RNAs (e.g., RNase I) or double stranded RNAs (e.g., RNase V1) on *in vitro* folded RNAs. These enzymes are too big to enter cells and thus RNA needed to be extracted and experiments performed

in vivo. The major disadvantage of these techniques was the lack of *in vivo* data. Since, RNAs can adopt multiple structure, which can be also differ due to protein binding, the applications of these techniques were limited. This led to the second generation of structure probing methods that used small molecules that could permeate cells and modify nucleotides selectively (DMS-seq (Rouskin et al., 2014), icSHAPE (Spitale et al., 2015)). These chemicals could either modify the bases or the backbone of nucleotides and the reactivity depended on the flexibility of nucleotides. For instance, a nucleotide participating in a base-pairing interaction would be much less flexible than the one that is single stranded. Thus, by measuring the reactivity at each nucleotide position, inferences could be made about the nucleotide status (free or interacting). Even though these methods provided information regarding RNAs *in vivo*, they do inform us of the exact interactions (e.g., base-pair) a particular nucleotide of an RNA was involved in.

To capture exact *in vivo* base-pairing interactions in RNAs, a technique called CLASH (Cross-Linking and Sequencing of Hybrids), was developed. CLASH specifically captured inter-RNA interactions mediated by specific proteins. (snoRNA- rRNA (Kudla et al., 2011) and miRNA-mRNA interactions (Helwak et al., 2013)) by ligating regions of RNAs that are involved in base-pairing (proximity ligation). Proximity ligation has been used to identifying higher order structure of DNA in techniques like chromatin conformation capture (Hi-C; (Lieberman-Aiden et al., 2009)) and thus provided an established approach to capture these interactions even for RNA. Similar approaches were developed to capture

unbiased, transcriptome-wide base-pairing interactions (Aw et al., 2016; Lu et al., 2016; Sharma et al., 2016). These methods combined proximity ligations with psoralen cross-linking, RNA base-pair intercalating agent, to specifically capture base-pairing interactions in an unbiased manner since they did not perform an IP for any protein. However, base-pairing interactions are not the only interactions important for RNA function. Structural changes in RNAs due to binding of RBP or RBP complexes play an equally important role. Thus, to capture all types of interactions a technique called RNA Proximity Ligation (RPL) was developed which cut and ligated all RNAs in whole cell extracts of yeast and HEK293 cells (Ramani et al., 2015) without enrichment of base-pairs. However, since most of the RNA inside the cell is non-coding, like rRNA and tRNA (Palazzo and Lee, 2015), very few of the RPL data mapped to the less abundant mRNAs. Thus, the lack of reads mapping to mRNAs in RPL indicated the necessity of enriching mRNPs before proximity ligations in understanding mRNP structure. Nevertheless, this demonstrated as in the case of chromatin, proximity ligations can be used to capture higher order RNA organization.

Thus, we developed a novel biochemical approach, RIPPLiT, that used the exon-junction complex as a handle to enrich for mRNAs and relatively less abundant ncRNAs like XIST. We combined, RIPiT which enriches for RNPs with proximity ligations for EJC associated RNAs, to specifically probe for higher order structure of mRNPs.

2. Materials and Reagents

2.1 Propagation and induction of stable cell lines

Cell growth and propagation reagents

- *Standard growth medium:* Dulbecco's modified eagle medium (DMEM; Life technologies, 11965-092) supplemented with 10% fetal bovine serum (FBS; Sigma, F2442-500ML) and 1% penicillin/streptomycin (Life technologies, 15140-122)
- *Tetracyclin*
- Trypsin-EDTA (Life technologies, 25300-062)
- Phosphate buffered saline (PBS)
- Harringtonine (LKT Laboratories # H0169-5mg)

2.2. Cell lysis and FLAG IP

- *Hypotonic Lysis Buffer (HLB):* 20 mM Tris-HCl pH 7.5, 15 mM NaCl, 10 mM EDTA, 0.5% NP-40, 0.1% Triton X-100, 1X Protease Inhibitor Cocktail (Sigma Aldrich # 4693159001), Harringtonine
- *Denaturing lysis buffer (DLB):* HLB supplemented with 0.1 % SDS and 0.1% sodium deoxycholate
- Branson Digital Sonifier-250 with microtip (Fisher, 15-338-125)
- Anti-FLAG agarose (Sigma, A2220)

- *Isotonic wash buffer (IsoWB):* 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.1% NP-40
- *Wash buffer 300 (WB300):* 20 mM Tris-HCl pH 7.5, 300 mM NaCl, 0.1% NP-40
- *Denaturing wash buffer (DWB):* IsoWB supplemented with 0.1 % SDS and 0.1% sodium deoxycholate

2.3. RNase T1 and FLAG elution

- RNase T1 (1000 U/uL) (Life Technologies # EN0541)
- Thermomixer (Eppendorf, 5355 000.011)
- FLAG peptide (Sigma, F3290) – prepare 5 mg/ml stock in Tris-buffered saline (TBS) and freeze aliquots at −20°C.

2.4. Second IP

- *2X IP2 buffer:* 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 20 mM EDTA, 1 % NP-40, 0.2% Triton X-100, X Protease Inhibitor Cocktail ()
- Protein-A Dynabeads (Life technologies, 10001D)
- Protein-G Dynabeads (Life technologies, 10003D)
- *Clear sample buffer (CSB):* 100 mM Tris-HCl pH 6.8, 4% SDS, 10 mM EDTA, 100 mM DTT

2.5 Dephosphorylation and phosphorylation

- polynucleotide kinase (PNK) wash buffer: 1X PNK (buffer 70 mM Tris-HCl, 10 mM MgCl2, no DTT), 5 mM DTT, 0.5% NP40
- Dephosphorylation Reaction Mix: 1X PNK buffer, 2.5 U T4 PNK (NEB), 1 mM DTT
- Phosphorylation Mix: 1X PNK buffer, 1 mM ATP, 1 U T4 PNK (NEB)

2.7 Ligation

- Ligation Mix: 1X T4 RNA ligase buffer, 1 mM ATP, 50 U T4 RNA ligase 1 (NEB, M0204)

2.8. RNA extraction

- Phenol:Chloroform:Iso-amyl alcohol, pH 4.5 (Life technologies, AM9720)
- *RNA precipitation stocks:* 5 mg/ml Glycogen (Life technologies, AM9510), 3M Sodium acetate, pH 5.2, 2M MgCl2, 200 proof (100%) ethanol

2.9. Estimation of ligation

- Hot Phosphorylation Mix: 1X PNK buffer, 150 µCi ɣ-^{32}P-ATP, 1 U T4 PNK
- *5'-end labeling:* T4 Polynucleotide Kinase (NEB, M0201S), 20mM DTT, 1mM ATP
- *2X formamide loading buffer (FLB):* 95% Formamide (deionized; Life technologies, AM9342), 10 mM EDTA, 0.025% Xylene cyanol, 0.025%

bromophenol blue; Alternative: Gel load buffer II (Life technologies, AM8547)

- *Gel electrophoresis equipment:* 20 × 27 cm Glass plates, Vertical gel electrophoresis apparatus, Heat shield
- Denaturing PAGE stocks:
    - Acrylamide mix: 26% acrylamide, 6M urea, 0.5X TBE
    - Dilution mix: 6M urea, 0.5X TBE
    - TEMED
    - 10% APS – Dissolve 1 g in 10 ml sterile water. Store at 4°C.
- *Oligos and ladders:* Low molecular weight ssDNA ladder (USB, 76410), 50 bp dsDNA ladder (Life Technologies)
- *Gel drying and imaging:* 3M whatman filter paper, Saran wrap, Gel dryer, Phosphorimager screens, Typhoon scanner

3. RIPPLiT strategy

RIPPLiT is comprised of two main steps, immuno-enrichment of the desired RNP complex through two-step IP followed by proximity ligation while the complexes are immobilized on solid support. For the two steps of immuno-enrichment to work, one of the proteins being IPed needs to be epitope tagged. Having the 2 steps aids in reducing the noise associated with non-specific pull down and also aids in pulling down multi-protein complexes if desired. We chose

to use the EJC as it is deposited on pol II transcribed, spliceosome spliced RNAs and remains tightly bound till the first round of translation. We have previously demonstrated that RIPiT protocol can be applied for the faithful purification of spliced but untranslated mRNPs (Singh et al., 2012; Singh et al., 2014). Further, similar to the RIPiT approach, RIPPLiT for EJC was also performed under native conditions. Data for EJC RIPiT demonstrated that the EJC footprints under native and cross-linking conditions are remarkably similar (Singh et al., 2012). To confirm there is no re-assortment of complexes during the RIPiT protocol, cell line mixing experiments were performed where extracts cell with Myc-tagged EJC were mixed with extracts from FLAG-tagged EJCs. Upon performing the RIPiT purification, we did not observe any Myc-tagged EJCs being pulled down with FLAG-tagged EJCs or vice versa. However, crosslinking significantly affected the yield of the purified sample. Hence, we performed RIPPLiT under native conditions.

By pulling down first with FLAG-Magoh followed by eIF4AIII, we enrich for RNAs associated with the EJCs and not individual proteins. Since, the EJCs are removed on first round of translation, we treat the cells with Harringtonine which blocks ribosomes at translation initiation, before cells extracts were made. Thus, we enriched for EJC associated RNAs that have most of their EJC intact. After the first IP step, complexes are mildly digested with RNase T1 which cuts RNAs after Guanosine instead of treating them with RNase I (as done for RIPiT). The aim for RIPiT was to obtain the *in vivo* footprint of EJCs and thus it required digestion of any region of RNA not bound by EJCs. However, in the case of RIPPLiT which

used T4 RNA ligase, it needed at least 3 nt overhangs on the substrate (Zhuang et al., 2012). Thus, by mild digestion and not cutting after every residue, we make sure there are sufficient overhangs for ligation. The complexes were then eluted under non-denaturing conditions, to keep the whole mRNP complex intact using FLAG-peptide. The second IP was done with antibodies linked to Protein-A dynabeads. Dynabeads are much smaller in diameter than sepharose beads (used for first IP) and thus help is reducing the required volume for following enzymatic reactions. Initial chromosome conformation capture experiments performed the ligations under high dilution conditions to ensure low non-physiological inter-complex interactions. However, since out complexes were immobilized on beads and the observation that EJCs do no re-assort in test tubes, we argued that we do not need to perform our enzymatic reactions under dilution conditions which allowed and potentially reduced the losses associated with dilution steps. Further, performing all enzymatic reactions while complexes were immobilized on solid support (dynabeads), also allowed us to easily change buffers and add new reaction mixtures. Upon treatment with RNase T1, RNAs are left with 3′ P and 5′ OH while T4 RNA ligase I (T4 Rnl I) requires the opposite. Hence, we need to first dephosphorylate the 3’ ends and then phosphorylate the 5′ ends of RNAs which can be done utilizing T4 Polynucleotide kinase (PNK). First in absence of ATP, it removes 3′ phosphates and then in presence of ATP, it adds phosphates to 5′ ends. This end-repair step also let us to incorporate radioactively label 5′-Ps and thus allowed us to perform 5′-end protection assay to demonstrate that the

observed shift in RNA sizes were indeed due to ligation step. Next, we ligate the ends of RNAs that are close together in space using T4 Rnl I by incubating the immobilized complexes at 25 °C overnight. All enzymatic reactions were done in 2 mL (round bottom) eppendorf tubes incubated on a thermomixer with shaking (1,000 rpm). This does not allow the beads to sediment and thus potentially reducing nonspecific interactions.

Although the RIPPLiT protocol was performed under native conditions, for other unstable RBPs, crosslinking steps can be incorporated as was done with RIPiT (Ricci et al., 2014; Singh et al., 2014). Thus, allowing for capturing mRNA protein interaction mediated by different types of RBPs.

RIPPLiT protocol

A. Cell culture and translational inhibition

1. Seed 6 X 10^6 HEK293 TREx cells on 150 mm plates in standard growth media.
2. Induce cells with Tet-HCl – 25 ng/ml for FLAG-Magoh, 10 ng/ml for FLAG-eIF4AIII for ~16-20 hours. This concentration was optimized to obtain near endogenous level of FLAG-Magoh expression (Singh et al., 2012).
3. Inhibit translation by adding 25 µL of Harringtonine (2 µg / mL) for 1 hr at 37 °C and thus enrich for EJC-associated mRNAs.

B. Preparation of cell extracts and 2-step immunoprecipitation

1. Discard media and rinse once with 10 mL PBS. Scrape the cells off in the remaining mount of PBS.
2. Re-suspend cells into PBS by vigorous pipetting.
3. Collect suspension and pellet cells at 400 x g for 10 min at 4 °C.
4. Re-suspend in ice-cold 3 mL hypotonic lysis buffer [Use 15 ml conical tubes for 3 mL size] per plate. Therefore, 3 X 6 = 18 mL needed here. Since, sonication could be done in a maximum of 4 mL volume for our particular probe, I re-suspended the cells in 3 mL hypotonic lysis buffer + PIC (50 X stock, so added 60 µL, 20 µL/ml) per 3 X plates. For instance, for 6 x 15 cm plates, make 2 batches for sonication of 4 mL each.
5. Incubate the samples on ice for 10 min after the addition of the lysis buffer.
6. Sonicate the extract at 40% amplitude using a Microtip for 6-seconds/ plate in bursts of 2 seconds with at least 10-second intervals for a total of 24 sec.
7. Add NaCl to 150 mM final concentration.
8. Spin the samples at 15,000 x g, 4 °C, 10 min. Bring up volume to 18 mL (3 times the number of plates) with hypotonic lysis buffer + 1X PIC.
9. Transfer 50 µL supernatant to a labeled tube to analyze total RNA or protein. Store at -20 °C.
10. Transfer remainder of supernatant to tube containing 250 µL/ plate anti-FLAG agarose beads, which have already been washed twice with 10 ml IsoWB (Isotonic wash buffer). Nutate at 4 °C for 1.5 hr.

11. After pelleting the beads, save 50 µL of supernatant and to check efficiency of depletion of the target protein. Store it at -20 °C
12. Wash beads 4 times with 10 mL IsoWB (ice-cold).
13. After 4th wash, transfer beads to 1.5 mL eppendorf by re-suspending beads in 1 mL IsoWB. Use low binding tubes when transferring beads to eppendorfs. Use 1 tube / 3 plates.
14. Distribute the IP into 2 parts and Add 125 µL / plate of isoWB+1:100 (Units/volume) RNase T1.
15. Incubate with intermittent shaking at 37 °C for 10 min.
16. Wash beads 4 times with 10 mL IsoWB (ice-cold) by transferring the beads back to a 15 mL falcon tube.
17. Remove beads in 2 X 2 mL eppendorf tubes with round bottom and add IsoWB, 125 X 3 = 375 µL of IsoWB + 250 µg/ mL FLAG peptide.
18. Elute by gentle shaking at 4 °C for 1.5 hr.
19. Remove 140 µL of elution and to the beads add 900 µL/ tube (2 tubes) of Isotonic lysis buffer. Mix to re-suspend and spin for 1 min at 400 X g at 4 °C.
20. Pool 860 µL supernatant with the elution in previous step. Take 50 µL out as "FLAG IP" fraction.
21. While the first IP is being eluted, wash 35 µL (for each IP from a 15 cm plate) of protein A dynabeads (for rabbit). Wash them twice with 1 mL of

PBST. Then add 400 µL of PBST and add 18 µL of anti-eIF4AIII antibody (Bethyl A302-980A) to the tubes.

22. Nutate at room temperature for 1 hr. Capture the beads on magnet and remove the supernatant.
23. Incubate the FLAG IP eluate with the eIF4AIII-conjugated beads, half for each fraction.
24. Nutate at 4 °C for 2 hrs. Keep the samples on magnet and collect the supernatant in a different tube as eIF4AIII IP depletion.
25. Wash beads 10 times with IsoWB (ice-cold). For each wash capture Dynabeads on magnet and remove the buffer. Add fresh buffer and completely re-suspend the beads. Change the tubes after 4th wash.

Since the RNPs were treated with RNase, they have a 3′ phosphate and a 5′ hydroxyl. For ligation to work, the ends need to be repaired to obtain 5′ phosphate and a 3′ hydroxyl.

Enzymatic reactions

All enzymatic reactions were performed in 2 mL round bottom eppendorfs since the reactions mix better in a 2 mL eppendorf than a 1.5 mL on a thermomixer.

1. Wash 3 times with isoWB and once with PNK wash buffer.

The enzymatic reaction didn't scale up in our hands. Hence, we decided to scale-down and carry out the reactions in batches of 3 plates.

2. Dephosphorylation reactions:

   Distribute the immunoprecipitated extracts into 2 tubes and add the following reaction mixture,

10X PNK buffer*	5 µL	1X, -DTT -NP40
RNA (beads)	~12 µL	
T4 PNK (10U/ µL)	2.5 µL	
20 mM DTT	2.5 µL	1 mM
Water	To 50 uL	

3. Incubate at 37 °C for 1 hr.

4. 5′ end phosphorylation reaction:

   For the 5′ end protection assay, we first end labeled the RNAs associated with mRNPs. For the hot phosphorylation reaction add following to the above reaction,

ɣ-32P ATP (150 µCi/ µL)	1µL
T4 PNK	1 µL

5. Incubate at 37 °C for 20 min.
6. This is followed by phosphorylation with cold ATP.

10X PNK buffer*	1 µL	
10 mM ATP (cold)	5.5 µL	~1 mM final

7. Incubate at 37 °C for 20 min.
8. Alternatively, for experiments without end labeling, 5’ end phosphorylation is done with only cold ATP with the following reaction mixture,

10X PNK* buffer	1 µL	
10 mM ATP (cold)	5.5 µL	~1 mM final
T4 PNK (10U / µL)	1 µL	

9. Incubate at 37 °C for 1 h.
10. Wash 3 times with isoWB and once with PNK wash buffer.
11. Ligation reaction:

    Distribute each PNK reaction into 2 tubes, + and - ligase.

    Ligation reaction mix (50 µL)

- Ligase			+ Ligase	
T4 Rnl buffer	5 µL		T4 Rnl buffer	5 µL
10 mM ATP	5 µL	1 mM final	10 mM ATP	5 µL
RNA	~12 µL		RNA	~8 µL
T4 RNA ligase 1	0 µL		T4 RNA ligase 1	5 µL
Water	To 50 µL		Water	To 50 µL

12. Incubate at 25 °C for overnight.
13. Save the supernatant of the ligation reaction (for both + and - Ligase).
14. To elution mRNP complexes, consolidate + and - ligase tubes.
15. Wash 3 times with IsoWB.
16. After the last wash add 36 µL of Clear sample buffer + 4 µL of 1 M DTT. Incubate at RT for 5 min, and then heat at 95 °C for 4 min, keep mixing the tubes occasionally.
17. Capture the beads on magnet and transfer all the supernatant (elution) to a new tube. Save 2 µL (from each + and - ligase tubes) as eIF4AIII IP for analysis of proteins on Western Blots and proceed to RNA extraction from the remaining.
18. Bring up the volume of the eluted sample to 300 µL with water. Add 1 µL glycogen + 30 µL 3 M NaOAc.

19. Add equal volume Acid (pH 4.5) phenol: chloroform: isaoamyl alcohol and vortex for 30 sec. Spin on tabletop centrifuge for 5 min. Take the aqueous phase (top) in a new tube.
20. Add equal volume PCIA and repeat twice followed by 2 chloroform extraction.
21. To the final 250 µL of aqueous phase + 650 µL 100% ethanol. Incubate at -20 °C overnight.
22. The next day, spin the tubes at 12000 X g, 4 °C, 1 hr.
23. Remove supernatant and add 1 mL 70% ethanol. Spin again for 10 min at 12000 X g at 4 °C.
24. Remove supernatant and let the pellet air dry.
25. Re-suspend the samples in 20 µL water. This can then be used for visualization of ligated products using denaturing gel electrophoresis (**Figure 3.1**), bioanalyzer (see **Figure 2.1B** for representative trace) and for preparing deep sequencing libraries.

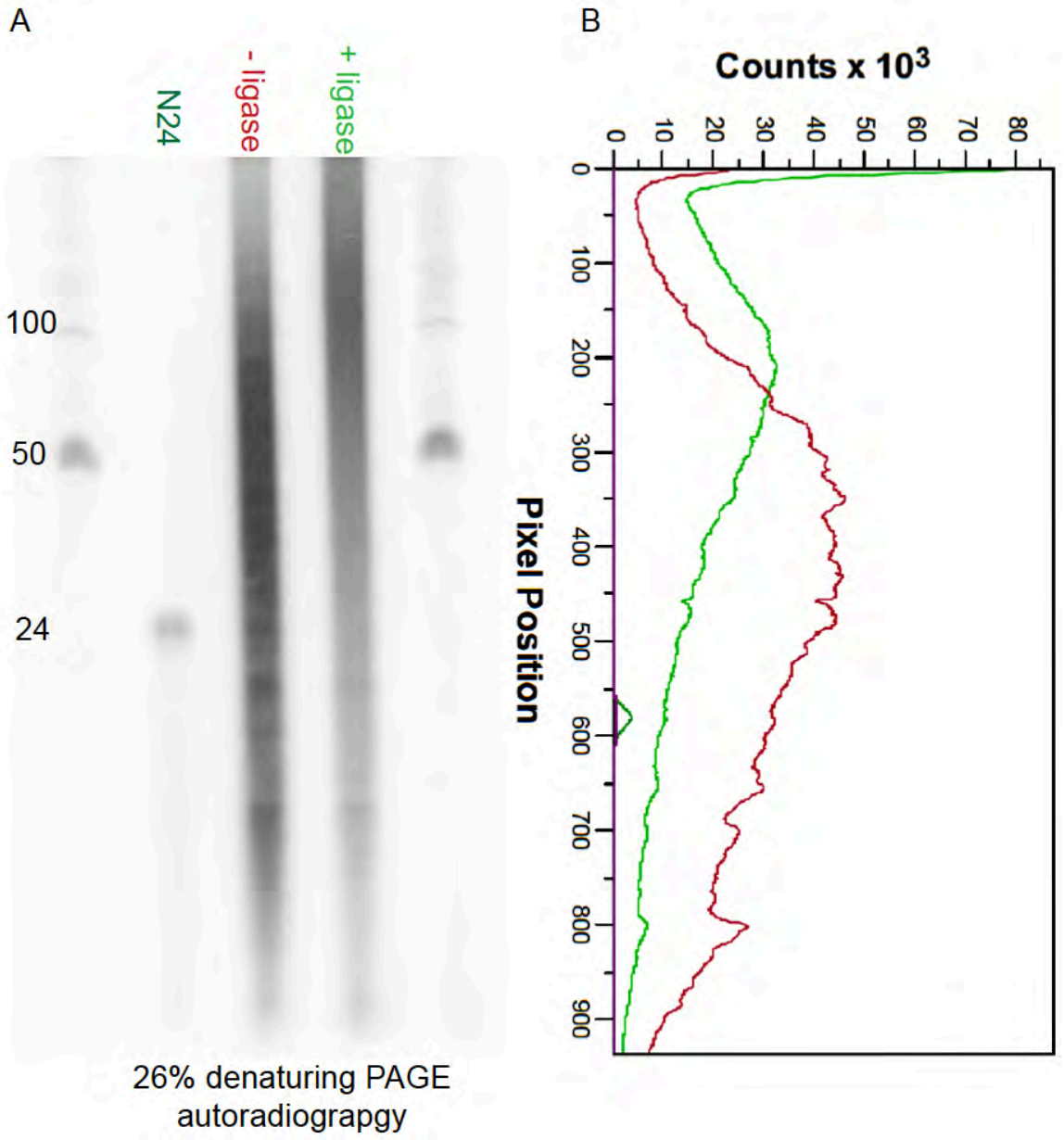


Figure 3.1. Shift in RNA size after ligation as visualized by Urea-PAGE and autoradiography

(A) 5′ end labeled (γ-^{32}P) RNA profiles for N24 (24 nucleotide oligo) and RIPPLiT experiments: + and - ligase. Autoradiogram of denaturing Urea-PAGE gel (26%) exposed for 1 hr.

(B) Signal quantified at each position across each lane using ImageQuant.

5’ end protection assay

We used a 5′-end label protection assay to test if the observed shift is RNA sizes was indeed due to the addition of ligase. We first end labelled the 5′ ends of RNAs before adding the ligase during the phosphorylation step of the protocol. Following this we distributed the IPs into 2 parts and added T4 Rnl I to one reaction but not to the other. Both reactions were then incubated overnight and RNAs were extracted the following day. Each of these reactions were further distributed into 2 tubes and either treated or not treatment with CIP. The rationale being if 2 RNA fragments were indeed ligated together, the labelled phosphates will be protected from CIP and thus can be visualized on a denaturing Urea-PAGE gel (**Figure 2.2**).

Other considerations

T4 PNK buffer contains 5 mM DTT which could affect the stability of antibodies holding the complexes together on the beads. DTT concentrations above 1 mM increase the risk of destabilizing the antibody and thus the IP of the complex (https://www.thermofisher.com/us/en/home/references/protocols/proteins-expression-isolation-and-analysis/antibody-protocol/dynabeads-co-immunoprecipitation-kit.html). Hence, we titrated the amount of DTT on a test oligo. A 20 nt nucleotide oligo was phosphorylated with T4 PNK enzyme for 1 hr and buffer containing increasing amounts of DTT (**Figure 3.2**).

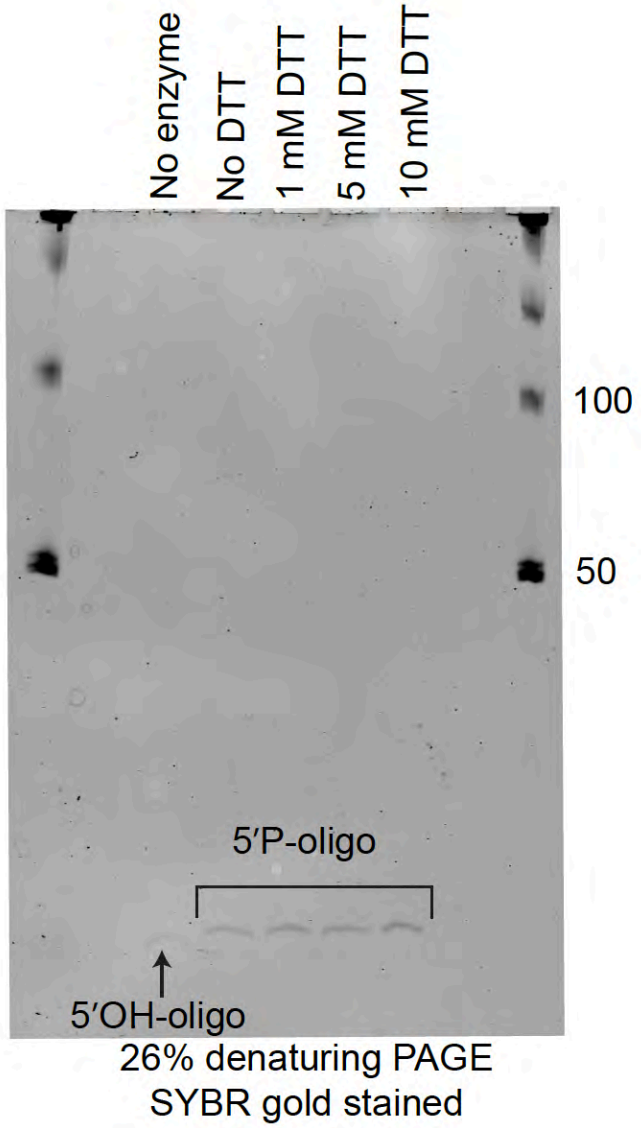


Figure 3.2. Concentration of DTT does not affect PNK phosphorylation reaction

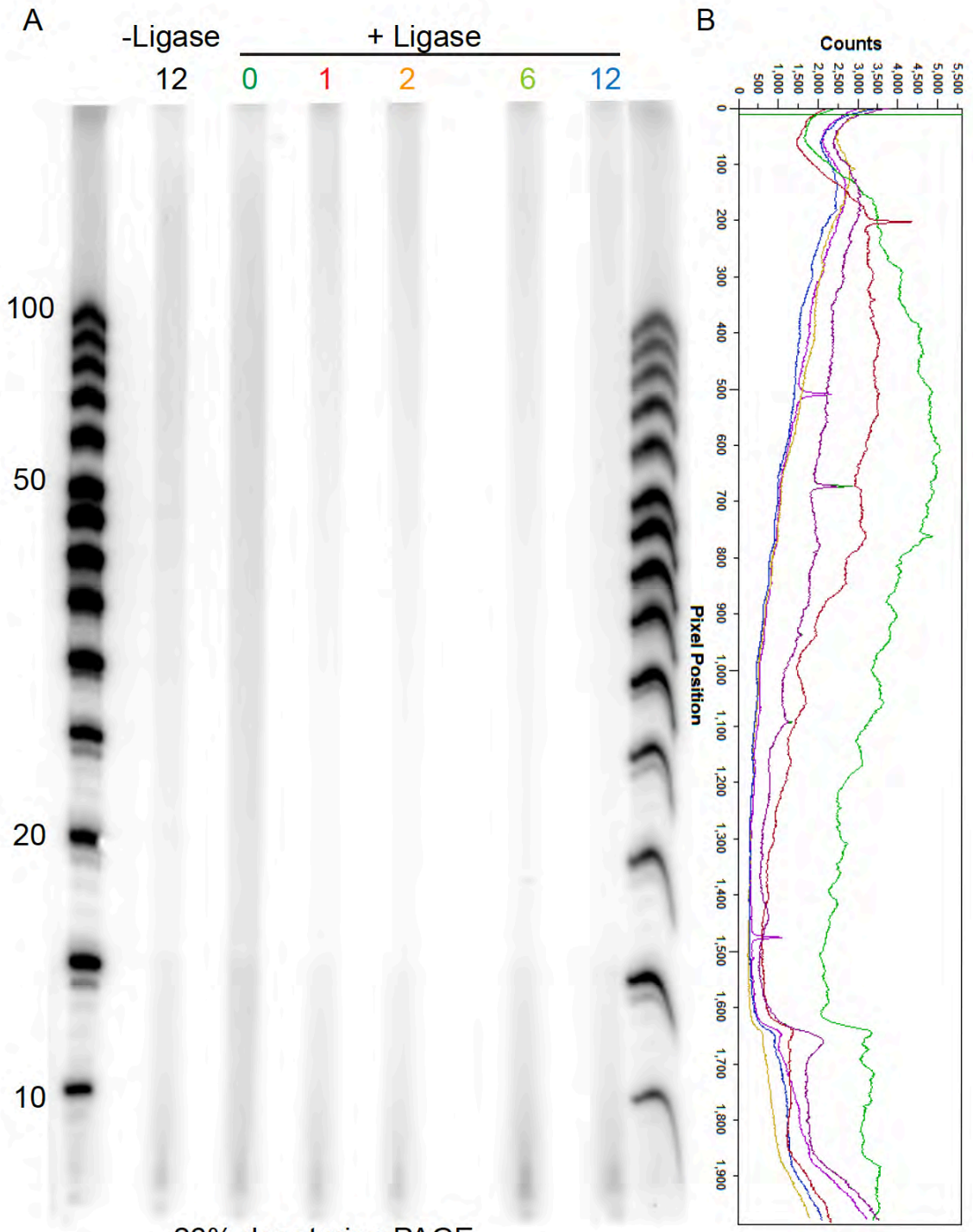
10 pmoles of 20 nucleotide RNA oligo (with 5′ OH) was phosphorylated using T4 PNK with varying concentrations of DTT. The RNA was then run of a denaturing gel (Urea-PAGE) and stained with SYBR gold. The 5′OH- and 5′P-oligos are labelled.

Phosphorylated products for all concentrations were compared with the recommended amount of DTT (5 mM). Interestingly, even PNK reactions without any DTT in the buffer yielded near complete phosphorylation, since we could not detect the 5′OH oligo. Since antibodies can tolerate 1 mM DTT, we used 1 mM concentration in our wash buffers as well as the PNK reactions.

Ligation time titration

Another important parameter that can potentially be changed is the time of incubation with ligase enzyme. For EJC RIPPLiT we performed a time course to follow the product formation, from 1 hr to 12 hr (**Figure 3.3**). We performed the EJC RIPPLiT with incorporation of hot ATP during the phosphorylation step. When run on a denaturing Urea-PAGE, we observed the product formation as early as 1 hr.

However, there might be a bias due to accessibility and sequence of the RNA fragments being ligated. Hence, we allowed the ligations to occur overnight for our experiments. Nevertheless, this demonstrated that for complexes not as stable as the EJCs, ligations could be performed for a much shorter time.



26% denaturing PAGE
autoradiograpgy

Figure 3.3. Ligation products for EJC RIPPLiT start to appear from 1 hour

(A) 5′ end labeled (γ-^{32}P) RNA profiles for RIPPLiT experiments with ligation performed for different durations. - ligase reaction was incubated for 12 hrs. Autoradiogram of denaturing Urea-PAGE gel (26%) exposed for 1 hr.

(B) Signal quantified at each pixel across each lane using ImageQuant.

RNA library size and length selection

Bionalayzer trace of RNAs extracted from RIPPLiT experiments showed that the RNA length distribution was 200-1,000 nt (**Figure 2.1B**). To capture the exact ligation junctions, it was imperative to obtain the longest read length and to make biologically significant conclusions, it was important to obtain sufficient depth of sequencing. The longest sequencing length, 600 nt (300 paired-end) can be obtained on an Illumina Miseq machine. However, it only provides ~40 million reads per run. On the other hand, we could obtain 500 nt (250 paired-end) with 10-fold more depth. Hence, we decided to sequence the RIPPLiT libraries using Nextseq 500 kit. Thus, RNAs were size selected for a range of 200-400 nt such that both read 1 and 2 would always overlap and no junction would be missed.

ChimeraTie strategy

Sanger sequencing of libraries revealed that the chimeric reads obtained through RIPPLiT consist of multiple (2 or more) fragments ligated together. Further, these fragments within a read did not follow the same order as that in reference **(Figure 2.1C)**. Hence, there was a need to develop a novel approach to capture these interactions present in chimeric reads. To this end, we developed, ChimeraTie, to map, analyze and visualize the chimeric reads.

Since, we did not know a priori the exact chimeric junction within a read, it was not possible to use any of the global alignment tools that force an alignment to span the entire length of the queried read (e.g., Tophat (Trapnell et al., 2009),

RSEM (Li and Dewey, 2011)). Thus, we required a tool that could perform local alignment which can identify regions within a read that map to a certain reference (Bowtie2 (Langmead and Salzberg, 2012), BWA-MEM (Li and Durbin, 2009), Segemehl (Hoffmann et al., 2014), STAR (Dobin and Gingeras, 2015)). Using a test dataset, we applied many of these tools to extract chimeric junctions. However, we faced multiple issues, some of which are listed below-

1. **Bowtie2**

   With the options, "--local -k N", Bowtie2 will use local alignment and report N distinct, valid alignments for each fragment (N is integer). The alignments are reported in the decreasing order of alignment score and a secondary alignment tag is added. A disadvantage of this approach is that multiple, high N (LIGR-seq found N = 50 optimal (Sharma et al., 2016)) values need to be tried which could make the program slower with increase in sequencing depth. Secondly, alignment search in not in any order, so it does not guarantee that the best alignment will be reported within the N options reported. Thus, there needs to be a balance for the value of N such that N alignments might contain the best alignment but without sacrificing time.

2. BWA-MEM

We also used BWA-MEM, a local alignment tool, with the “-a” option. With this option in place, it provides all alignments it finds for a given read which are flagged secondary alignment. However, a major difference of BWA-MEM vs Bowtie2 was that some of the secondary alignments reported by BWA-MEM overlapped (**Figure 3.4**). Thus, this would pose an additional issue when identifying chimeric junctions. Further, “-a” option requires BWA-MEM to report all alignments for all the input reads. This affects the alignment time inversely, especially as the read length increases.

Figure 3.4. Sample read with overlapping secondary alignments

(A) Sample alignment for one read with multiple secondary alignments reported by BWA-MEM. Read ID: red, remaining fields: black for each alignment.

(B) CIGAR strings (left) for the 4 alignments from (A) for the same read describing how the read aligned (5′ to 3′) to reference. Numbers indicate number of nucleotides and Letters describe the operation performed on the number of nucleotides preceding it. H (Hark clipped): nucleotides removed from the reported alignment, M (Matches): matches present in alignment, D (Deletions): nucleotides absent in alignment but present in reference, S (soft clipped):

nucleotides present in the read but not aligned. Colored bars (right), scaled to read length, illustrate the CIGARs for each alignment. Blue: hard or soft clipped, yellow: matches and red: deletions. Areas within the green dotted lines show the overlapping matches reported by BWA-MEM.

---

3. Segemehl

Another tool that can perform mapping of fragments within a read is Segemehl. By providing the option “-S”, allows Segemehl to perform local alignment read splitting. As it first performs global alignment on all the reads and then local alignment on unmapped reads, it much more efficient with respect to time. However, when same test data was mapped with either Segemehl or Bowtie2, we obtained more alignments with Segemehl than Bowtie2 (**Figure 3.5**). The reads mapped by Bowtie2 were a subset of the ones mapped by Segemehl. However, the remaining alignments provided by Segemehl were all secondary alignments and sometimes not even within the same gene. Thus, Segemehl appeared to have similar issues as BWA-MEM as well as higher possibly false positive alignments compared to Bowtie2.

Segemehl Alignment

```
M02096:114:000000000-AEH6V:1:1113:23019:22034 1:N:0:1   0       ENSG00000108848|ENST00000505658|LUC7L3|17|7040|ENSE00003673736|50746542|50746702|1
|9    1       255     51M     *       0       0AAGTAGAGATCGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAAAACA      IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIII     NM:i:0  MD:Z:51 NH:i:1  XI:i:0  XL:i:5  XA:Z:Q  XX:i:43 XY:i:93 XQ:i:1  XP:Z:ENSG00000108848|ENST00000505658|LUC7L3|17|7040|ENSE0000
3673736|50746542|50746702|1|9 XU:i:51 XS:i:64 XC:Z:ENSG00000108848|ENST00000505658|LUC7L3|17|7040|ENSE00003673736|50746542|50746702|1|9       XV:i
:1    XT:i:32
M02096:114:000000000-AEH6V:1:1113:23019:22034 1:N:0:1   0       ENSG00000108848|ENST00000505658|LUC7L3|17|7040|ENSE00003673736|50746542|50746702|1
|9    1       255     51M     *       0       0AAGTAGAGATCGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAAAACA      IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIICIIIIIII
IIIIIIIII     NM:i:0  MD:Z:51 NH:i:1  XI:i:0  XL:i:5  XA:Z:Q  XX:i:95 XY:i:145        XQ:i:2  XP:Z:ENSG00000108848|ENST00000505658|LUC7L3|17|7040|
ENSE00003673736|50746542|50746702|1|9 XU:i:51 XS:i:64 XC:Z:ENSG00000108848|ENST00000505658|LUC7L3|17|7040|ENSE00003673736|50746542|50746702|1|9XV:
i:1   XT:i:32
M02096:114:000000000-AEH6V:1:1113:23019:22034 1:N:0:1   0       ENSG00000108848|ENST00000505658|LUC7L3|17|7040|ENSE00003673736|50746542|50746702|1
|9    1       255     51M     *       0       0AAGTAGAGATCGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAAAACA      IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIII     NM:i:0  MD:Z:51 NH:i:1  XI:i:0  XL:i:5  XA:Z:Q  XX:i:147        XY:i:197        XQ:i:3    XP:Z:ENSG00000108848|ENST00000505658|LUC7L
3|17|7040|ENSE00003673736|50746542|50746702|1|9        XU:i:51 XS:i:64 XC:Z:ENSG00000108848|ENST00000505658|LUC7L3|17|7040|ENSE00003673736|50746542
|50746702|1|9 XV:i:1  XT:i:32
M02096:114:000000000-AEH6V:1:1113:23019:22034 1:N:0:1   0       ENSG00000108848|ENST00000505658|LUC7L3|17|7040|ENSE00003673736|50746542|50746702|1
|9    1       255     51M     *       0       0AAGTAGAGATCGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAAAACA      IIIIIIIIIIIIIIIIIIIIIIIIIIIIIGIIIIIIIIIIII
IIIIIIIII     NM:i:0  MD:Z:51 NH:i:1  XI:i:0  XL:i:5  XA:Z:Q  XX:i:199        XY:i:249        XQ:i:4    XP:Z:ENSG00000108848|ENST00000505658|LUC7L
3|17|7040|ENSE00003673736|50746542|50746702|1|9        XU:i:51 XS:i:64
M02096:114:000000000-AEH6V:1:1113:23019:22034 1:N:0:1   0       ENSG00000108848|ENST00000505658|LUC7L3|17|7040|ENSE00003673736|50746542|50746702|1
|9    11      255     41M     *       0       0CGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAAAACA        CCIIIIIIIIIGGIIIGIIIGIGIIIIGGIIIIIIIGIIIII        NM
:i:0  MD:Z:41 NH:i:1  XI:i:0  XL:i:5  XA:Z:Q  XX:i:1  XY:i:41 XQ:i:0  XC:Z:ENSG00000108848|ENST00000505658|LUC7L3|17|7040|ENSE00003673736|50746542
|50746702|1|9 XV:i:1  XT:i:32
```

Bowtie2 Alignment

```
M02096:114:000000000-AEH6V:1:1113:23019:22034   0       ENSG00000108848|ENST00000505658|LUC7L3|17|7040|ENSE00003673736|50746542|50746702|1|9    1
       2       198S51M20S      *       0       0CGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAAAACAGAAGTAGAGATCGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAAAACAGAAGTA
GAGATCGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAAAACAGAAGTAGAGATCGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAAAACAGAAGTAGAGATCGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAA
AACAGAAGTAGAGATCGACGAAGA       CCIIIIIIIIIGGIIIGIIIGIGIIIIGGIIIIIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIICIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIII     AS:i:51 XS:i:41 XN:i:0  XM:i:0  XO:i:0    XG:i:0  NM:i:0  MD:Z:51 YT:Z:UU
M02096:114:000000000-AEH6V:1:1113:23019:22034   256     ENSG00000108848|ENST00000505658|LUC7L3|17|7040|ENSE00003673736|50746542|50746702|1|9    11
       2       41M228S *       0       0       CGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAAAACAGAAGTAGAGATCGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAAAACAGAAGTAG
AGATCGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAAAACAGAAGTAGAGATCGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAAAACAGAAGTAGAGATCGACGAAGAAGCAGAAGCCATGATCGATCAGAAAGAAA
ACAGAAGTAGAGATCGACGAAGA       CCIIIIIIIIIGGIIIGIIIGIGIIIIGGIIIIIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIICIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIII     AS:i:41 XS:i:41 XN:i:0  XM:i:0  XO:i:0    XG:i:0  NM:i:0  MD:Z:41 YT:Z:UU
```

Figure 3.5. Segemehl reports more alignments compared to Bowtie2

Same fastq file was used to map to the same reference by either Segemehl or Bowtie2. Image displays reported alignments, 5 for Segemehl and 2 for Bowtie2, for the same sample read. For each alignment, Read ID is in red while the remaining fields are in black. Note: Only the last 2 alignments reported by Segemehl are also reported by Bowtie2.

4. STAR

STAR has capabilities to detect chimeric junctions within input reads. However, it assumes a maximum of one junction within a read (Dobin and Gingeras, 2015); also see STAR manual) presumably because it is designed to map splice junctions coming from either alternative or circular splicing. However, based on our initial Sanger sequencing we had identified multiple reads with more than 2 fragments ligated together. Thus, using STAR we would miss the junctions present in these reads.

Even though Bowtie2 performed the best amongst all the tools we tested, it still missed many of the expected alignments in our test data. Further, "-k" option made mapping significantly slower with no guarantee for the reporting of the best match for a given read. This made it imperative to develop a novel approach, which we call ChimeraTie, to best capture chimeric junctions.

In ChimeraTie, instead of reporting multiple secondary alignments for reads in a fastq file, we first report the best alignment for all reads in local alignment mode. The remaining unmapped parts of each read are put in a new fastq file and Bowtie2 is called again to map this new file. This process continues till either all of the fragments within a read are mapped or the fragments are smaller than the minimum alignment length provided to Bowtie2 (set to 16 nt by default). While reporting each alignment, ChimeraTie specific tags (see below) are added to identify the exact location of the fragment within a read, length of alignment, alignment span in reference and also the iteration in which this fragment was

mapped to retain all the important information for detailed analysis of the dataset (e.g., length distribution of alignments, order of fragments within the read, order within the reference).

Analysis pipeline used for RIPPLiT

Upon sequencing these libraries, the first analysis step involved merging the reads 1 and 2 based on their overlap into single reads. We performed this using Paired-End AssembleR (pear, v0.9.6) (Zhang et al.,2014) with its default options and any read pair without an overlap was discarded.

pear -f file_R1.fastq -r file_R2.fastq -o file_R1R2_merged.fastq (I)

Next step involved removal sequencing adapters which was achieved using cutadapt v 1.7.1.

cutadapt -n 2 -g adapter_1_seq -a adapter_2_seq
-o outfile.fastq infile.fastq (II)

-n COUNT

Try to remove adapters at most COUNT times. (Default = 1)

-g ADAPTER, --front=ADAPTER

Sequence of an adapter that was ligated to the 5' end.

-a ADAPTER, --adapter=ADAPTER

Sequence of an adapter that was ligated to the 3' end.

In our case we used SMARTer-seq smRNA kit to make deep sequencing libraries. This adds 3 nt on the 5′ end and ~15 nt polyA tail on the 3′ end of reads that need to be trimmed. Length of the polyA tail was empirically determined. This step was also performed using cutadapt.

cutadapt -u 3 -u -15 -o infile.fastq outfile.fastq (III)

-u LENGTH, --cut=LENGTH

Remove bases from the beginning or end of each read.

If LENGTH is positive, the bases are removed from the beginning of each read.

If LENGTH is negative, the bases are removed from the end of each read.

For the purpose of our analysis, we locally generated a transcriptome by mapping the - ligase control to GrCh37 assembly using RSEM (bowtie2 and default parameters) (Li and Dewey, 2011); One of the output files of RSEM provided expression levels for each isoform of all the genes present in the GTF file. Using this, we chose the highest abundance isoform to create a reference with single transcripts for each gene.

EJCs are deposited on RNAs by the spliceosome and removed by the ribosome. Hence, repeat RNAs, rRNA and snRNAs, need to be filtered out from EJC associated deep sequencing libraries. The PEAR merged reads, after adapter removal and end trimming, were first mapped to rRNAs using ChimeraTie followed

by mapping to snRNAs. The unmapped reads obtained after these steps were then mapped to the locally-generated transcriptome.

```
python PATH/chimera-tie.py -v -f infile.fastq -x bt2_index \
--bopts '--local -D 20 -R 3 -N 0 -L 20 -i S,1,0.50 -p10 --rdg 5,6 --rfg 5,6'       (IV)

-f FASTQ, --fastq FASTQ
                    Input fastq file - assumes single sided (default:
                    None)
  -x BOWTIE2_IDX_PREFIX, --bowtie_idx BOWTIE2_IDX_PREFIX
                        Path to bowtie2 idx file (default: None)
  --bopts BOWTIE2_OPTIONS, --bopts BOWTIE2_OPTIONS
                    Optional bowtie2 alignment options (default: --local
                    -D 20 -R 3 -N 0 -L 16 -i S,1,0.50)
  -v, --verbose
                  Increase verbosity (specify multiple times for more)
```

ChimeraTie adds the following optional tags that are specific to it,

XQ:i:	Iteration of mapping in which this fragment was mapped
XX:i:	Start position of alignment in read
XY:i:	End position of alignment in read
ZM:i:	Match length of alignment
ZR:i:	Length of the read this alignment came from
ZS:i:	Span length of alignment in reference

ChimeraTie uses Bowtie2 in local alignment mode to iteratively map fragments within a read. Hence, some ligation that span short distances can be misidentified as gaps in a single alignment or a gap can be misidentified as two different fragments ligated together. To minimize this issue, we modified Bowtie2's read and reference gap to empirically determine the best options for penalties (**Figure 3.6**). Bowtie2 in local alignment, identifies the best alignment by adding bonuses for each match and subtracting penalties for each difference (mismatch, gap, etc). Hence, the length of gap allowed is a function of the alignment length and as the aligned sequence length increases, the weight of the penalty decreases. Therefore, it is not possible to completely eliminate the false positives chimeric junctions or false negative singleton reads.

These read and reference gap penalties can be specified using the options-

--rdg <int1>,<int2>

Sets the read gap open (<int1>) and extend (<int2>) penalties.

A read gap of length N gets a penalty of <int1> + N * <int2>.

Default: 5, 3.

--rfg <int1>,<int2>

Sets the reference gap open (<int1>) and extend (<int2>) penalties.

A reference gap of length N gets a penalty of <int1> + N * <int2>.

Default: 5, 3.

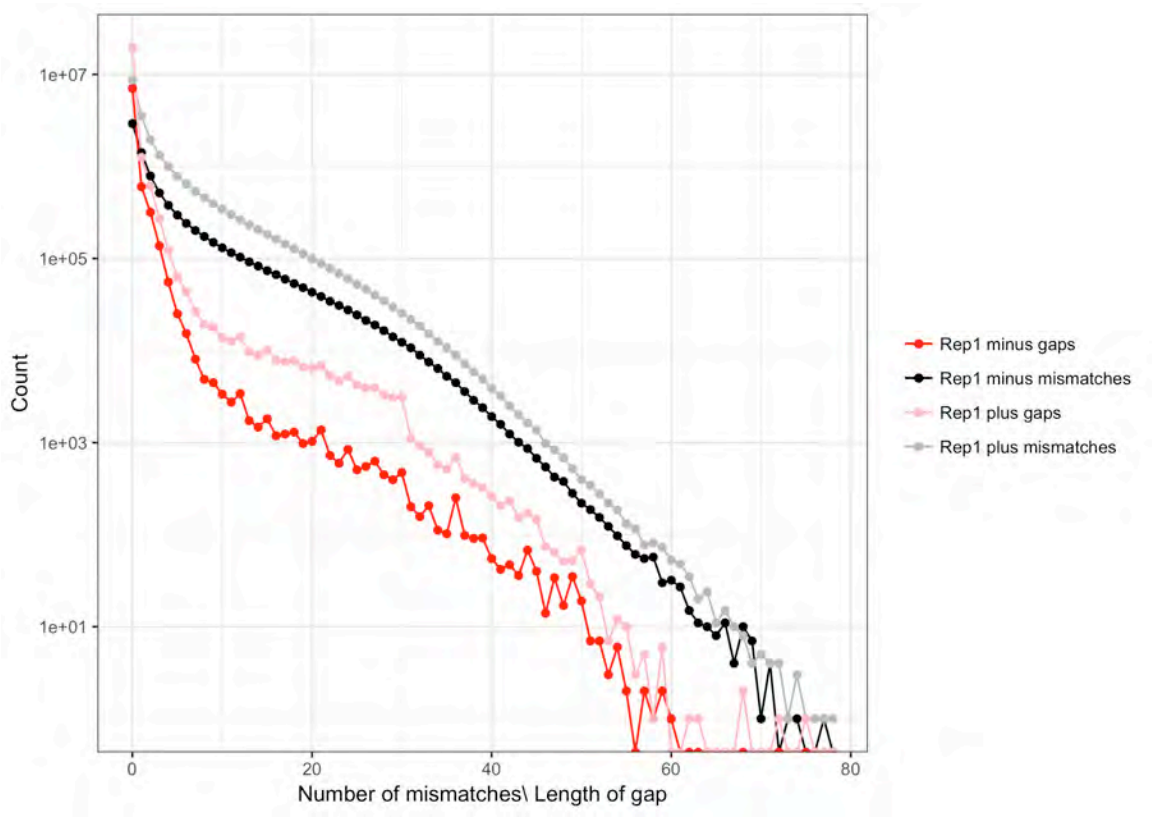


Figure 3.6. Bowtie2 local alignments on RIPPLiT long reads had high frequency of mismatches and gaps

Data is shown for replicate 1 minus and plus ligase libraries. X-axis shows either number of mismatches (black and gray curves) or length of gaps (red and pink curves) within a reported alignment. Y axis displays the counts for each value of X.

We increased the stringency of our mapped alignments by first removing multi-mapping reads (tag XS). A disadvantage of having long reads that are mapped with local alignment is that they tend to have longer stretches of mismatches and gaps with an increase in alignment length. We filtered the alignments containing more than 3 gaps (tag XG) or 3 mismatches (tag XM) to gain higher confidence in our alignment results.

```
samtools view infile.bam | grep -v "XS:i" | awk '($2 != 4)' | \
cat bam.header - |  samtools view -bS - > outfile.bam
```
(V)

```
bamtools filter -tag 'XM':'<=3' -in infile.bam -out outfile.bam
```
(VI)

```
bamtools filter -tag 'XG':'<=3' -in infile.bam -out outfile.bam
```
(VII)

The reads remaining after the filtering steps were converted into a pairwise interaction file. This script takes the bam file as input and first classifies reads into 2 categories- “S” or “Singletons”: reads with no chimeric junctions and chimeric reads. The chimeric reads are further classified into “D” or “Direct” junctions in which the 2 chimeric fragments were present without any intervening nucleotide sequence within the read or “I” or “indirect” where the 2 chimeric fragments had an intervening sequence (>0) between them. Each chimeric junction is then reported

as a pair with a “D” or an “I” tag while the singleton reads are reported as paired with themselves with a “S” tag.

python bam2itx_strand.py -b infile.bam -v (VIII)

The output pairwise interaction file that have the extension “.itx” consist of the following tab separated fields-

Field	Description	Example
JuncType	Type of chimeric junction. “S”, “I” or “D”	D
QNAME	Query name	NS500602:302: HY73KBGXY:1:22311: 14925:17267
RNAME_frag1	Reference name for first fragment.	RNA18S5
POS_frag1	1-based leftmost mapping Position for first fragment.	1
FLAG_frag1	Combination of bitwise FLAGs for first fragment. Same as in sam file.	0
STRAND_frag1	Strand inferred from sam FLAG for first fragment.	+
XX_frag1	Start position in read of first fragment, 0 based.	XX:i:0
XY_frag1	End position in read of first fragment, 0 based.	XY:i:17
XQ_frag1	Iteration of mapping in which this fragment was mapped for first fragment	XQ:i:1
ZM_frag1	Match length of alignment for first fragment	ZM:i:17

ZS_frag1	Span of alignment in reference for first fragment	ZS:i:17
ZR_frag1	Length of the read this alignment is part for first fragment	ZR:i:17
RNAME_frag2	Reference name for second fragment.	RNA18S5
POS_frag2	1-based leftmost mapping Position for second fragment.	1486
FLAG_frag2	Combination of bitwise FLAGs for second fragment. Same as in sam file.	0
STRAND_frag2	Strand inferred from sam FLAG for second fragment.	+
XX_frag2	Start position in read of second fragment, 0 based.	XX:i:17
XY_frag2	End position in read of second fragment, 0 based.	XY:i:93
XQ_frag2	Iteration of mapping in which this fragment was mapped for second fragment, 0 based.	XQ:i:0
ZM_frag2	Match length of alignment for second fragment.	ZM:i:76
ZS_frag2	Span of alignment in reference for second fragment.	ZS:i:76
ZR_frag2	Length of the read this alignment is part for second fragment.	ZR:i:97

The pairwise interaction file is compared to a genome annotation file to annotate the pairs reported therein. The step was designed for cases where the mapping is first performed on a genome reference and then in step the transcript

annotating for the genomic location can be added. However, we mapped to the transcriptome for EJC RIPPLiT data and so mapping to the genome followed by comparison to the transcriptome has not been fully tested.

```
python itx2subset_strand.py -i infile.itx -g annotation.gff -s F \
-o outfile__annot.itx                                                    (IX)
```

```
-i ITX_FILE, --itx ITX_FILE
                    Input, sorted, itx file - output of bam2itx.py
                    (default: None)
  -o OUTPUT_ITX_FILE, --annotated_itx OUTPUT_ITX_FILE
                    Itx file with annotation. Gene names, start, end,
                    strand added (default: None)
  -g GENE_ANNOTATION, --gene_annotation GENE_ANNOTATION
                    Path to gene annoation GFF file (default: None)
  -s (F,R,N), --strandness (F,R,N)
                    Strandness of the RNA-Seq Library. The fragments are
                    going to be matched to the gff entries according to
                    this parameter. If you do not want to consider strand,
                    set it to NF: Forward, R: Reverse, N: None (default:
                    None)
```

This step appends transcript annotation to the interaction file for each fragment and also a list of headers from the sam file. We used a modified GFF annotation file for our analysis that followed the format given below,

> Sample tab separated gff entry with header.

#bin chrom processed entry_type txStart txEnd strand name exonic_part_number name2

1 RNA5S1 None Edited 1 122 + RNA5S1 0 5S

2 RNA28S5 None Edited 1 5071 + RNA28S5 0 28S

The pairwise interaction file obtained was then converted to a matrix for a given set of mRNAs. Currently, there is a limit to the number of rows (bins) in matrix that can be handled which is set to 50,000. Thus, matrix at 1 nt resolution for 50 Kb or 1 gene per bin for 50,000 inter-RNA interactions can be visualized. For the analysis of EJC RIPPLiT data, we filtered chimeric junctions within reads that are directly next to each other in the read (“Direct” junctions).

python itx2matrix_hakan.py -i overlapped_itx_infile.itx -r 'gene_name' --bsize n -D -o outfile.matrix.gz (X)

-i ITX_FILE, --itx ITX_FILE

Input chimeraTie ITX file (default: None)

-o OUTPUT_FILE

Output matrix file (default: None)

-r REGIONS, --regions REGIONS

Regions to be picked from the interaction file (default: All)

--bsize

BIN_SIZE

-S

Include singletons in matrix (default: False)

-I

Include indirect itx in the matrix (default: False)

-D

Include direct itx in the matrix (default: False)

-g GENOME

Genome Assembly (default: hg19)

-v, --verbose

Increase verbosity (specify multiple times for more)

The matrices thus obtained can be converted into heatmaps using the heatmap.pl script that is part of Hi-C cWorld suit (github).

perl heatmap.pl -i infile.matrix.gz -v (XI)

Discussion

Here we have described a novel biochemical, RIPPLiT, and bioinformatics, ChimeraTie, approach to capture the higher order interactions of mRNAs associated with the RNA-binding proteins(s) of interest. A major advantage of both the tools is their universal applicability. RIPPLiT can be applied to any RNP of interest to capture both base-pairing (e.g., base-pairs bound by Stau1 (Ricci et al., 2014)) or RNA-folding mediated by associated proteins (e.g., EJC). Although we performed EJC RIPPLiT under native conditions, for less stable complexes like Stau1 interactions, formaldehyde crosslinking step could easily be incorporated in the protocol. In fact, we have applied RIPiT for capturing Stau1 footprints (Singh et al., 2014) under formaldehyde crosslinking conditions. This protocol could be easily combined with proximity ligations as done in EJC RIPPLiT. Similarly, we developed ChimeraTie to map EJC mediated RNA interactions. However, since it maps any type of chimeric junction within a read, it could easily be repurposed to capture novel alternative splice junctions, junctions formed by circular RNAs or even fusion transcripts.

Even though both of these are powerful approaches, they have some limitations. Specifically for EJC RIPPLiT, we cannot identify the origin of the captured junction- whether it was an a base-pairing interaction that brought the 2 regions together or was the interaction an effect of RNA folding. Further, since the ligations are not performed *in situ,* we cannot completely rule out the possibility of non-physiological interactions happening as an effect of the biochemical

purification. EJCs have been shown to not re-assort during various biochemical purification and immunoprecipitation steps. However, this needs to be tested when applying to other complexes that are not as stable as the EJC.

Similarly, ChimeraTie is one the many new tools developed for mapping chimeric reads obtained as a result of proximity ligations. There need to be a concerted effort to test and benchmark all the tools to identify the best approach in handling such complicate datasets.

Overall, RIPPLiT and ChimeraTie, together provide a novel and powerful toolkit to elucidate some of the most fundamental questions of RNA biology.

Endnotes for Chapter III

Chapter III is adapted from the following:

Metkar M., Ozadam H., Lajoie B. R., Dekker D., Moore M. J., RIPPLiT and ChimeraTie: High Throughput Tools for Understanding higher order RNP structures. (manuscript in preparation- to be submitted to Methods for review)

Author Contributions

M.M. and M.J.M. originally conceived the project, with M.M. executing all wet bench experiments. B.R.L. and H.O. conceived and wrote ChimeraTie and M.M. implemented all computational analyses with input from M.J.M and J.D. All authors contributed to data analysis and interpretation. With input from all authors, M.M., J.D. and M.J.M. were primarily responsible for writing the paper.

Chapter IV

Discussion

DISCUSSION

An early observation in the field of DNA was that some DNA fragments migrate slower than others. This led to the hypothesis that DNA has an inherent curvature which creates more friction and thus slows migration through a porous gel (Marini et al., 1982). To prove this was indeed the case, Ulanovsky *et al.*, incubated oligos of known sequence with T4 DNA ligase and visualized the products using 2D-gel electrophoresis to identify DNA circularization (Kotlarz et al., 1986). 2nd dimension was run with an DNA intercalating agent, Chloroquine, with the assumption that it would distort linear and circular DNAs differently which is what they observed. Further, using denaturing gels, they demonstrated that these circular DNA species migrated slower than their linear counterparts. This concept, that curvature in DNA could be capture by ligating the ends of DNA that are close together in space (proximity ligation) was later used to study the bending of DNA mediated by proteins (Kotlarz et al., 1986). They argued that if a protein bends the DNA, it should affect the frequency at which these ligations happen. This was further helped with the identification of condition under which intra-molecular ligations were favored over inter-molecular. Thus, they applied proximity ligation to a known DNA fragment with and without a protein binding to it and visualized the products on a denaturing acrylamide gel. With this, they were able to demonstrate that more circular DNA mono-molecules were created in the presence of the protein compared to without. However, these techniques could not

be applied to elucidate DNA bending and architecture in a high-throughput way, since no technology existed to study multiple DNA fragments at the same time.

The advent of PCR provided an opportunity to probe for multiple regions at the same time to understand 3D genome organization. Dekker *et al.* were the first to combine proximity ligations with semi-quantitative PCR to understand the 3D conformation of *Saccharomyces cerevisiae* chromosome III and were able to show that it organizes as a contorted ring (Dekker et al., 2002). Further improvement in technologies involving PCR and high-throughput sequencing, allowed for the interrogation of 3D interactions with higher resolution (e.g., promoter to gene vs whole chromosome) (de Wit and de Laat, 2012). Though these techniques were high-throughput and unbiased compared to other microscopy based (e.g., FISH) techniques available at the time, the interaction quantification was limited by the choice of PCR primers used. Capturing interactions by applying proximity ligation was further revolutionized with the introduction of Hi-C which incorporated biotin at the junctions between 2 fragments, thus allowing for enrichment of only ligated DNA pieces without the need for region specific PCR primers (Lieberman-Aiden et al., 2009). Thus, this allowed for the genome-wide survey of all interactions in the same experiment in an unbiased manner. Further improvements in Hi-C technique and decreasing sequencing costs, allowed to study 3D interactions within chromatin in different biological states (e.g., cell cycle phases; (Gibcus et al., 2018)) to different organisms (e.g., *Plasmodium falciparum* (Teng et al., 2015)).

Proximity ligations combined with NGS were first used for RNA in a technique called CLASH (Cross-linking, Ligation, And Sequencing of Hybrids) which was specifically designed to capture inter-RNA interactions (Kudla et al., 2011). The chimeric products, fragments of 2 different RNAs joined together, were in fact a coincidental discovery. The authors were interested in mapping the precise binding site of an RNA binding protein (RNA helicase prp43) on its target RNA, snR53 box C/D snRNAs using another technique called cross-linking and analysis of cDNAs (CRAC) (Kudla et al., 2011). The protocol for CRAC involved ligation of an oligonucleotide linker to enriched RNA fragments. During this step, at a very low frequency (0.46%), they identified snoRNAs fused to their targets, 18S rRNAs in the same read, thus leading to the development of CLASH.

An important lesson in this discovery was the significance of examining data in an unbiased manner and the effort to study the 0.46% reads that would have been normally discarded. This serendipitous discovery with RNA and the numerous important discoveries through Hi-C have led to the development of multiple other approaches combining Proximity Ligations, NGS and various other biochemical approaches including RIPPLiT.

EJC RIPPLiT was specifically designed to capture the higher order confirmations of pre-translational mRNPs. To this end we first enriched EJC associated mRNPs, thus mRNAs that are spliced but not yet translated and captured interactions mediated by both base-pairing and RNA folding due to EJC (and its associated proteins) using proximity ligations. These events were then

deconvoluted using NGS. Further, due to bioinformatics challenges (like absence of a mapping tool that can map reads with multiple fragments within it), we developed a new tool called ChimeraTie to map, analyze and visualize chimeric reads. Together the biochemical approach, the bioinformatics tools and polymer analysis, helped us delineate the fundamental rule of mRNP packaging. Analyzing data for more than 400 mRNAs we were able to show that they are linearly organized and tightly packed into a flexible rod like structure, irrespective of their length when mRNAs are packaged into mRNPs.

Overall RNA proximity techniques can be classified into 2 main categories and 2 subcategories within them. The main categories are- 1) Unbiased ligations of all RNAs within cells (IP independent), and 2) Ligating RNAs bound to a specific protein or a protein complex (IP dependent). These can be further classified as a) probing for specifically base-pairing interactions, or b) all interactions mediated by RNA folding (through base-pairing or protein binding) (**Table 4.1**, **Figure 4.1**). Table below classifies all the current approaches in each of the categories-

	IP dependent	IP independent
Base-pair	CLASH, hiCLIP	LIGR-seq, PARIS, SPLASH
Base-pairs + proteins	RIPPLiT, MARIO	RPL

Table 4.1. Classification of RNA proximity ligation techniques

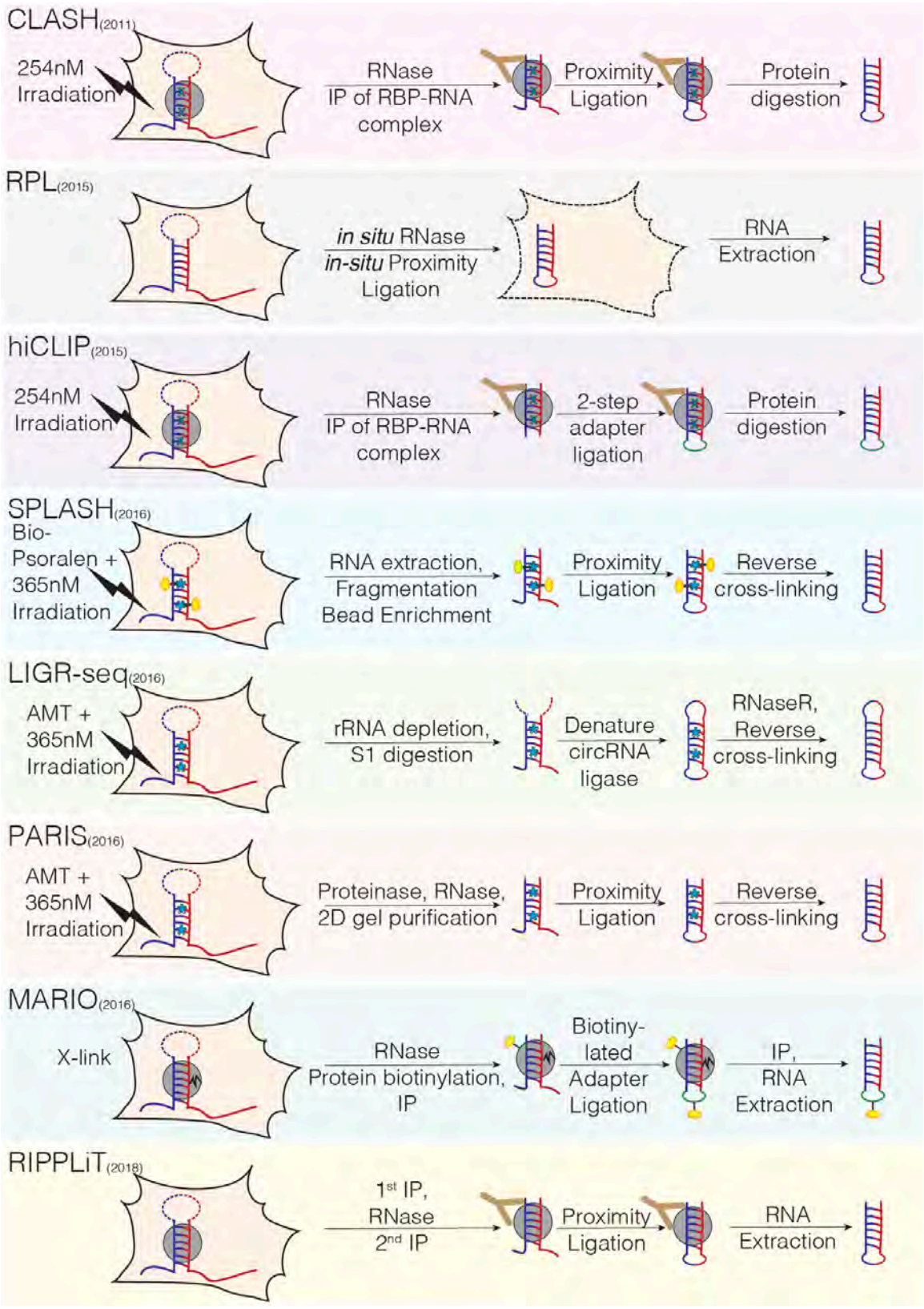

CLASH(2011)
254nM Irradiation
RNase IP of RBP-RNA complex
Proximity Ligation
Protein digestion
RPL(2015)
in situ RNase
in-situ Proximity Ligation
RNA Extraction
hiCLIP(2015)
254nM Irradiation
RNase IP of RBP-RNA complex
2-step adapter ligation
Protein digestion
SPLASH(2016)
Bio-Psoralen + 365nM Irradiation
RNA extraction, Fragmentation Bead Enrichment
Proximity Ligation
Reverse cross-linking
LIGR-seq(2016)
AMT + 365nM Irradiation
rRNA depletion, S1 digestion
Denature circRNA ligase
RNaseR, Reverse cross-linking
PARIS(2016)
AMT + 365nM Irradiation
Proteinase, RNase, 2D gel purification
Proximity Ligation
Reverse cross-linking
MARIO(2016)
X-link
RNase Protein biotinylation, IP
Biotiny-lated Adapter Ligation
IP, RNA Extraction
RIPPLiT(2018)
1st IP, RNase 2nd IP
Proximity Ligation
RNA Extraction

Figure 4.1. Schematic comparison of RIPPLiT with other RNA proximity ligation methods

Similar to **Figure 1.3**, but with RIPPLiT. Lines in red and blue are spatially proximal RNA fragments coming from either the same transcript or two different transcripts. Gray discs: RBPs, blue asterisk: AMT intercalated within RNA, blue cross: UV crosslinked bases, black wiggly lines: protein-RNA crosslink, yellow-ovoid disc: biotin, brown “Y”-shape: antibodies, green line: oligonucleotide adapter, nude structure: cell, lightening shape: UV irradiation.

Lessons from ncRNAs and the open questions

XIST

X inactive specific transcript (Xist) has been one of the best studied ncRNA due to its cellular abundance and its important biological function in dosage compensation in female mammals (Pintacuda et al., 2017). It does so by physically coating the X chromosome and also recruiting multiple protein complexes that aid in silencing (Ng et al., 2007; Pintacuda et al., 2017). However, the process of how XIST and its associated proteins mediate X chromosome silencing is still debated. Since, XIST structure is important for both coating the X chromosome and recruiting the silencing machinery, a lot of effort has been put in understanding the structure of XIST (Pintacuda et al., 2017). Using RIPPLiT we were able to add to this knowledge by capturing XIST's higher order organization.

Data from PARIS (Lu et al., 2016), technique that captured base-pairs, showed that there are 2 long hairpins in the first and the last exons of XIST. However, as PARIS captured direct base-pairs in an ensemble way and XIST contains multiple repeat regions, it could be possible that there are different regions of XIST that interact within its different molecules forming smaller hairpins. But these would appear as long hairpins in an ensemble study capturing only base-paired regions. On the other hand, RIPPLiT captured interactions at the base of the loops and so, the exact regions forming the base-pairs were not visualized by it **(Figure 2.9D)**.

Further data from techniques measuring nucleotide flexibility, SHAPE-MaP (Smola et al., 2016), have shown that repeat E of XIST is less flexible *in vivo* compared to *in vitro*. However, this data alone was insufficient deduce the reason for this reduced flexibility. In RIPPLiT data, we saw a strong signal in the same region. Given that this region doesn't contain stable hairpins, and that it has signal in RIPPLiT with low accessibility in SHAPE-MaP *in vivo*, implied that this region is tightly folded as a result of protein binding.

Thus, XIST exemplifies the effective use of complementary techniques in elucidating different aspects of the same question. One of the salient features of EJC RIPPLiT is its ability to capture not just the base-pairing interactions but also the interactions mediated by RNA folding through the proteins bound by it. However, it is not possible to differentiate the obtained signal based on its origin. Thus, to effectively use RIPPLiT to understand detailed 3D conformation of specific RNAs, it is essential to combine it with other structure probing techniques. Further, systematic deletion experiments revealed that different repeats of XIST play an important role in its function (Wutz et al., 2002). In this light, it will be interesting to study the changes in XIST 3D conformation upon deleting these regions using RIPPLiT (and other structure probing techniques) and how they affect its function. In the current study, RIPPLiT was applied to the EJC and thus, interactions mediated by other proteins like various hnRNP proteins (which are not known to interact with EJCs) were excluded. For instance, terminal exons of XIST where we observe only a limited number of chimeric junctions (**Figure 2.9B**) contains the

long hairpins and also extensive binding site for the nuclear binding protein hnRNP U (Yue and Ogawa, 2018). Similarly, other ncRNAs like NEAT and MALAT1 are de-enriched in EJC libraries due to a lack of EJCs. However, they could also contain specific structural features that help in their function. It would be useful to combine RIPPLiT with other RNA binding proteins (like hnRNP) to understand the higher order conformations of these ncRNAs.

rRNAs

Given the fact that ~90% of total RNA in the cell is ribosomal and the EJCs are removed by the ribosome, EJC RIPPLiT libraries consisted of ~27% rRNAs. Interestingly, the absolute number of 18S chimeras was larger than the absolute number of 28S chimeras (e.g., 79,597 vs 32,130 in Replicate 1; **Figures 2.5A** and **Figure 2.6A** inset numbers). This was largely due to 18S fragments being more abundant than 28S fragments in our samples (e.g., 3,096,819 vs 2,287,536; Table X), despite 18S rRNA being only 37% the length of 28S rRNA (1,869/5,070*100). One contributing factor was the almost complete absence of fragments from the parts of both rRNA molecules not modeled into the 3D structures (the white areas in Figures X and X), likely due to the loss of these regions upon RNase digestion. A larger percentage of 28S nucleotides is missing compared to 18S. Other factors possibly contributing to 18S rRNA enrichment are the temporary structural bridge between PYM and EJCs (EJCs are removed by PYM during the pioneering round

of translation (Gehring et al., 2009)) and/or the effect of harringtonine, which stalls translation at the stage of large ribosomal subunit addition (Fresno et al., 1977). It is worth noting, however, that examination of multiple genes with 5'UTR introns revealed no substantial difference in RIPPLiT coverage over the 5'UTR compared to the coding region. Thus, our samples do not appear to be dominated by mRNA molecules on which the small subunit has scanned through the 5'UTR, removing EJCs as it goes, and stalled at the start codon.

Since the ribosome scans the mRNA to decode it, it seemed plausible that EJC RIPPLiT captured chimeric junctions between the mRNA and mRNA binding regions of the ribosome. These interactions could potentially shed light on the features of rRNA-mRNA interactions during the first round of translation (e.g., are there differential interactions between different parts of an mRNA and the rRNA?). However, we observed no such region-specific interactions on either the 18S or the 28S rRNA and the most abundant mRNAs. One major factor contributing to this could be that we are looking for interactions between 1 rRNA vs >10,000 mRNAs. This issue could be solved by increasing the depth of our libraries. Thus, unravelling interactions specifically during the pioneering round of translation for either specific mRNAs or a set could be an interesting avenue to pursue.

Other small ncRNAs

In our RIPPLiT protocol, after RNA extraction, we size selected RNAs greater than 200 nt. This resulted in de-enrichment of other abundant smaller RNAs like tRNAs, snRNAs, snoRNAs, miRNAs etc. CLASH with Ago protein was able to identify *in vivo* binding sites for miRNAs on target mRNAs as well as classifying miRNA:mRNA interactions based on their binding pattern (Helwak et al., 2013). Similarly, LIGR-seq identified novel and unexpected interactions between snoRNAs:mRNAs (Sharma et al., 2016). Recently, a tRNAs, through tRNA derived small RNAs, have been shown to regulate multiple steps of an mRNA's life cycle (Li et al., 2018). Since, EJC RIPPLiT enriches for interactions for mRNAs, higher order complexes involving these smaller ncRNAs and mRNAs can also be captured. Thus, it might be interesting to size select RIPPLiT libraries for lengths less than 200 nt that might contain these kinds of inter-RNA (ncRNA:mRNA) interactions, providing a global snapshot of mRNA regulation through small ncRNAs.

What makes mRNAs different from ncRNAs?

The major difference we observed between chimeric junction pattern for >450 mRNAs versus ncRNAs (e.g., rRNAs, XIST, snRNAs) was the near homogeneous pattern of chimeric junctions as seen on a junction heatmap. Since enzymatic and structural ncRNAs like rRNAs and XIST perform functions that

require a conserved structure, the chimeric junction heatmaps for these showed locus-specific interactions. However, we observed no such specific patterns (locus-specific) interactions for mRNAs. This indicated that there are no conserved higher order structures within mRNAs and that molecules of a particular mRNP can vary slightly in structure from each other giving rise to a near homogenous chimeric junction heatmap. However, this variability in pattern could arise due to the difference in depth of chimeric junctions. For instance, + ligase replicate 1 RIPPLiT dataset contained ~50-fold more chimeric junctions for rRNA (18S rRNA: 79,597 chimeric junctions) vs mRNA (UBR4: 1,611 junctions). It could be argued that due to the 50-fold less chimeric junctions for mRNAs, locus-specific interactions were not observed. To test this hypothesis, we down sampled rRNA chimeric junctions and binned it similar to mRNAs. Since, UBR4 contained ~1,600 chimeric junctions in replicate 1 + ligase sample, we randomly selected 1,600 chimeric junctions for 18S and 28S rRNA and plotted a heatmap such that each bin represents 1% of the transcript. Even after down sampling rRNA chimeric junctions (18S and 28S rRNAs) we observed the locus specific interactions (**Figure 4.2**). Thus, demonstrating that the non-locus specific interaction pattern observed for mRNAs was not an artifact of lower chimeric junction depth.

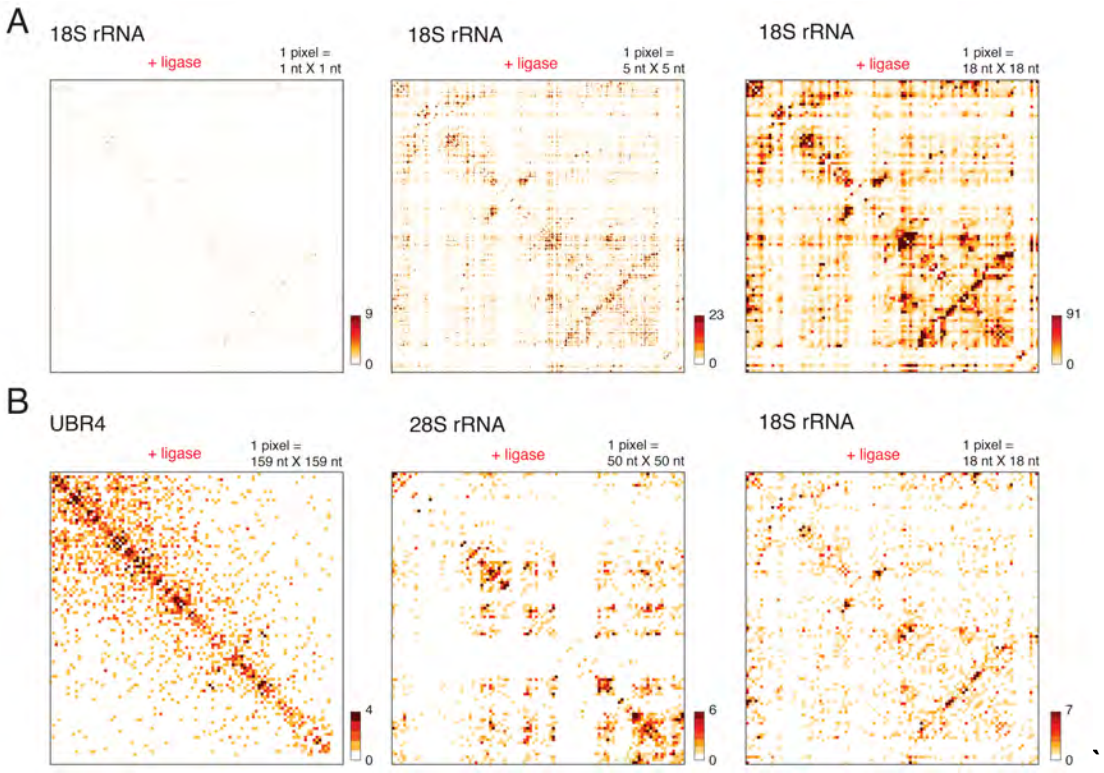


Figure 4.2 Differences in heatmap pattern persist even after down-sampling number of junctions in ncRNAs to match mRNAs

(A) Chimeric junction heatmaps for 18S rRNA (replicate 1 plus ligase). Data was aggregated at different bin size to show the effect of binning on the observed locus-specific interaction patterns. Color scale indicates the number of chimeric junctions.

(B) Same as in (A) but for UBR4 (mRNA; 1,611 chimeric junctions), and 28S and 18S rRNA (ncRNAs) downsampled to 1,600 chimeric junctions. Heatmaps were binned such that each bin is equal to 1% of the transcript.

Why were intramolecular interactions within mRNPs largely ignored in proximity ligation approaches?

One major difference between for proximity ligations when applied to DNA or RNA is that for DNA intra-chromosome interactions are more abundant than inter-chromosome (Lajoie et al., 2015). While for RNA, inter-RNA, especially ones involving ncRNAs (rRNAs, snoRNAs, snRNAs) are much more easily captured than intra-RNA. Further, many of these ncRNAs function through base-pairing interactions (e.g., miRNA:mRNA, rRNA:snoRNA) and thus these interactions are more abundant than inter-chromosomal interactions (since chromosomes for territories; (Lajoie et al., 2015)). Moreover, interactions in DNA are currently measured at 100s of kilobases vs for RNAs it is possible to point the exact nucleotides that are ligated (single nucleotide resolution).

Presumably, given the abundance of ncRNAs, their structures are relatively consistent from molecule to molecule (e.g., 18S rRNA will more or less fold in a similar way), and that they function through base-pairing with their targets, elucidating the base-pairing interactions were the focus of many of the RNA proximity ligation approaches (CLASH, hiCLIP, SPLASH, PARIS, LIGR-seq)(Helwak et al., 2013). Further, structure probing methods like DMS-seq (Rouskin et al., 2014) and icSHAPE (Spitale et al., 2015) observed that mRNAs were much more flexible *in vivo* compared to *in vitro* which led to the conclusion that mRNAs are largely unstructured within cells and thus would not provide specific and consistent interactions similar to rRNAs or snRNAs. In addition,

mRNAs are lot more diverse than any specific ncRNA. For example, a eukaryotic cell contains approximately in the order of 10^5 copies of snRNAs as well as mRNAs (Palazzo and Lee, 2015). However, these are distributed amongst ~10 types of snRNAs while 10,000s of distinct mRNAs. Thus, it is far easier to obtain sufficient data for a single type of snRNA compared any specific mRNA. This issue was further exacerbated by low percentage of chimeric reads obtained in the proximity ligation-based approaches (They typically range from 0.0006% to <5%; **Table 4.2**). Only exception being MARIO with 30% chimeric reads in one dataset. However, the control libraries contain ~7-8% chimeric reads (Nguyen et al., 2016).

Method	Cells	Number of reads	Number of Interactions	Percentage
CLASH	Yeast: Saccharomyces cerevisiae	25,000,000	115,000	0.46
	Human: HEK293	*Numbers not provided		<2
hiCLIP	Human: HEK293, high RNase	2,551,621	36,140	1.42
	low RNase	3,022,687	23,251	0.77
PARIS	Human: HEK293T-Rep3	56,829,056	1,199,910	2.11
	HeLa_Low_RNase	63,073,219	1,714,832	2.72
	HeLa_High_RNase	70,003,455	2,987,980	4.27
	Mouse: mES_1	28,405,622	790,987	2.78
SPLASH	Lymphoblastoid Cells Total RNA Replicate 1	53,747,987	311,079	0.58
	Lymphoblastoid Cells PolyA Replicate 1	183,913,864	160,800	0.09
	Human ES PolyA Replicate 1	159,412,735	73,407	0.05

	Human RA PolyA Replicate 1	153,884,298	77,245	0.05
LIGR-seq	HEK293T: Read1, + AMT, + ligase	171,239,817	1,011	0.0006
	Read2, - AMT, + ligase	258,157,264	4,111	0.002
RPL	Yeast: Saccharomyces cerevisiae	304,000,000	851,200	0.28
MARIO	Mouse: ES 1	45,702,794	13,848,413	30.30
	MEF	83,083,324	17,616,980	21.20
	Brain	36,463,565	2,877,233	7.89
RIPPLiT	HEK293- Rep1	48,977,271	1,297,588	2.65

Table 4.2. Number and percent of reads with chimeric junction for different RNA proximity ligation techniques

To alleviate these issues various approaches were employed (e.g., immunoprecipitation with RNPs; RIPPLiT, CLASH (Kudla et al., 2011), hiCLIP (Sugimoto et al., 2015), Ribozero Gold; LIGR-seq (Sharma et al., 2016), GeneRead rRNA Depletion Kit; MARIO (Nguyen et al., 2016)) to reduce the amount of rRNAs (~90% of all RNA inside the cells), from the libraries. However, given the abundance of rRNAs, it was impossible to get rid of it completely. On the contrary, since the 3D structures of rRNAs (Khatter et al., 2015) are available and their processing through snoRNAs is well studied, they were extensively used to validate many of the techniques. For instance, RIPPLiT, RPL and SPLASH used the decay in ligation frequency with increase in 3D distance. While CLASH, PARIS, LIGR-seq used known snoRNA:rRNA interactions to validate their techniques. Along with validation, these methods were able to identify novel interactions like snoRNA:mRNA interactions.

Potential method changes to increase the yield of chimeric junctions

As stated in the previous section and **Table 4.2** one of the major caveats for all the RNA proximity ligation methods is the low yield of chimeric reads with maximum yield being ~4% (except MARIO with 30% with very high noise; ~8%). This would most likely be the next important developmental phase in the evolution of these techniques. Some of the factors (see below) affecting this have been

explored in the different methods and maybe built upon for future method development,

1. RNase treatment: Since, the T4 RNA ligase I requires ssRNA ends (Zhuang et al., 2012) that are in close proximity to ligation the amount of digestion can be modulated to obtain optimal distance between ends. PARIS (Lu et al., 2016) performed experiments in HeLa cells with under 2 RNase conditions, low and high, and observed a ~40% increase (2.7-4.2%) in chimeric read yield in the more stringent digestion.

2. Ligase concentration and time: Another parameter that can possibly be varied is the concentration of ligase and the time of ligation. However, at least for RIPPLiT increasing ligase concentration did not seem to affect the ligation efficiency (data not shown). Many of the ligation protocols are done at 16-25 °C overnight under native or crosslinking conditions. Ligation time course for RIPPLiT showed that the reaction appeared to reach saturation after 8-12 h of ligation.

3. Oligonucleotide adapter: hiCLIP (Sugimoto et al., 2015) and MARIO (Nguyen et al., 2016) ligated an adapter in between the 2 fragments of spatially proximal RNAs. Potentially, this oligo can help in identifying junctions during sequencing or even enriching for chimeras if they are biotinylated (MARIO). However, this requires a 2-step ligation protocol with end repair (3′ end dephosphorylation and 5′ phosphorylation) and washes after each step. This could significantly

affect the final yield of chimeric RNAs and explain the low yield (hiCLIP) or high noise (MARIO).

4. Sequencing deeper and longer read lengths: Another way to increase the yield of chimeric reads would be a brute force approach to sequence the libraries deeper. Along with this, it will prove useful to have capabilities to sequence longer read lengths. Specifically for RIPPLiT, where we observed more than 2 RNA fragments ligated in single read, made it unavoidable for us to sequence the whole read so that we would not miss any junction. Thus, as sequencing technology improves, and we are able to sequence deeper with longer reads will no doubt help with higher number of chimeric reads.

For methods involving immunoprecipitation before proximity ligation, one major factor that seems to affect the yield of chimeric reads is the choice of protein. For instance, CLASH when applied to snoRNA binding proteins (Kudla et al., 2011) contained 0.46% chimeric reads, while when applied to Ago1 (Helwak et al., 2013) contained ~2% chimeric reads. One difference between the 2 experiments was that snoRNA CLASH was done with yeast while Ago1 was done with human cell lines. Thus, to test exactly how much the complex of interest plays a role or was this difference in chimeric reads only due to the cell type needs to be tested by performing the experiments in the same cells type simultaneously.

Future directions

Limitations and caveats of EJC RIPPLiT

As stated above the biggest limitation of RIPPLiT was the low yield (~2.5%; **Table 4.2**) of chimeric reads. Further, since RIPPLiT involves immunoprecipitation before performing the proximity ligations, the data is biased by the RNA fragments bound by the protein or the protein complex being pulled down. For instance, intronless genes are selected against upon IP with EJCs since EJCs are not deposited on them (see **Figure 2.3**). Also, since the ligations are not performed *in situ,* we cannot completely rule out the possibility of non-physiological interactions happening as an effect of the biochemical purification. Although EJCs have not been shown to re-assort during various biochemical purification and immunoprecipitation steps (Singh et al., 2012), RNP complex stability needs to be tested when applying to other complexes that are not as stable as the EJCs. Other caveats specifically for EJC RIPPLiT include- we can only obtain data for pre-translations RNPs as EJCs are removed during the first round of translation by the ribosome (Gehring et al., 2009). Further, we cannot identify the origin of the captured junction- whether it was a base-pairing interaction that brought the 2 regions together or was the interaction an effect of RNA folding.

Similarly, ChimeraTie is one the many new tools developed for mapping chimeric reads obtained as a result of proximity ligations. There need to be a

concerted effort to test and benchmark all the tools to identify the best approach in handling such complicate datasets.

Testing and expanding the proposed model for mRNP organization

Based on our analysis of the scaling plots, we speculate that mRNPs are organized as elongated structures inside cells. This conclusion is based upon the result that the chimeric junction frequency decreases gradually with increase in nucleotide distance. However, this model does not predict the positioning of the mRNAs inside the packaged RNPs. Multiple models could be envisioned for such an arrangement. For instance, mRNA is packaged around a proteinaceous core (as predicted for YBX-1 protein binding an mRNA (Skabkin et al., 2004)), or the proteins coat the mRNA such that it's on the inside of the mRNP complex (FGRY2 protein in *Xenopus* oocytes protects the bound RNA from digestion (Matsumoto et al., 2003)). Further, the mRNA could be arranged as sequential array of loops of nearly equal length, or the mRNA could be arranged as coils in these particles. All of these hypotheses along with confirming the model can be tested by combining experimental procedures and polymer simulations.

For example, the mitotic chromosomes were proposed to form an array of loops emanating from a backbone of condensin proteins (Gibcus et al., 2018). Using Hi-C experiments they obtained scaling plots with interactions frequencies plotted against nucleotide distances. This was then tested by simulating a polymer

arrangement where the polymer forms a loop array inside a cylindrical shape with an axial scaffold. Further, the model was validated by systematically knocking down scaffold proteins (here condensin I and II) and analyzing the change in chromosome shape, Hi-C contact frequencies as well as expected changes in scaling plots.

A similar approach could be envisioned for testing mRNP organization. However, this requires a prior knowledge of the proteins forming the backbone. Based on EJC footprinting experiments we know that the EJC and its associated proteins are an important part of mRNP organizational backbone (Sauliere et al., 2012; Singh et al., 2012). Thus, it would be interesting to systematically knock down either the EJC proteins or the SR and SR -like proteins and then measure the changes in chimeric junction frequencies. Knockdown of eIF4A3 caused a decrease in association of SRSF1 and SRSF3 (Singh et al., 2012). Thus, to begin with these 3 proteins- eIF4A3, SRSF1 and SRSF3 could be knocked down prior to performing RIPPLiT.

Validating RIPPLiT data

By combining principles of polymer physics with RIPPLiT, we were able to speculate that mRNAs are organized as flexible rod-like structures when they form mRNPs. Recently, 2 studies performed single molecule FISH (smFISH) experiments to study the relative distances between the 5′ and 3′ ends of specific

mRNAs by using probes against the mRNA ends (Adivarahan et al., 2017; Khong and Parker, 2018). Interestingly, they observed that in untranslated mRNAs, inside the nucleus or cytoplasm, the 5′ and 3′ ends are seldom together corroborating the conclusion that mRNPs form elongated structures. However, with smFISH it is not possible to visualize the complete mRNP particles. Also, only a few mRNAs can be studied in a given experiment. One way to visualize whole particles would be to purify the mRNPs and study them under electron microscopes (EM). Immunoprecipitated samples after the first or the second IP without RNase treatment could be isolated and visualized on EM grids. However, samples after the first IP (after non-denaturing elution) would include non-specific RNPs that have not undergone 2-step purification and thus would contain some impurities. On the other hand, samples after the second IP cannot be eluted under non-denaturing conditions and thus would need to be visualized with the attached antibodies. Another way to purify these mRNP particles would be to replace the second IP against eIF4A3 with IP against poly dT which can then be cleaved using RNase E (cleaves single stranded A and U nucleotides). These could then be visualized under EM to measure the shape and size of these particles. RNase treatment needs to be excluded between IPs for this approach as it could adversely affect the 3′ end mRNP structure. EJCs bind upstream of an exon-exon junction and therefore 3′ UTRs will be more sensitive to digestion. Since this approach requires the poly A tail as a handle for 2nd IP, it would affect the yield significantly. An alternative approach to purify EJC associate mRNPs would be to separate the

particles by density-gradient centrifugation after the 1st IP with one of the EJC proteins followed by their visualization under EM. A caveat however, of this approach is that the samples would need to be cleaned to remove the gradient material (e.g., sucrose) before being visualized under EM.

Overall EM would provide a good visual verification of the particle shape. However, it will not provide the information of the identity of the visualized mRNP since it will be an ensemble of mRNPs being purified and visualized.

One of the major limitations for the validation for RIPPLiT was the absence of an mRNP with known 3D structure, similar to rRNAs. It would be illuminating to purify a specific mRNP using methods like RNA-antisense purification followed by proximity ligation and visualization under EM. Using this approach, we will be able to characterize, in detail, organization of that particular mRNP at very high resolution as well as have a visual validation for the molecule.

Are mRNAs circular throughout their life cycle?

Electron microscopy was employed on two rat cell types- the somatotrope (anterior pituitary), source of growth hormone (or somatotropin), and the mammotrope, source of prolactin to study the organization mRNAs associated with ribosomes on the rough endoplasmic reticulum (Christensen et al., 1987). These membrane-associated polysomes were seen to form circular (~80%) with approximately 6-7 ribosomes present in each “circle” or “G” shaped (~20%)

structures with approximately 8-9 associated ribosomes. This observation led to the hypothesis that the ribosome can re-engage an mRNA upon completing one round of translation. This was further validated by reconstituting the interaction between cap binding protein, eIF4E, with the polyA binding protein Pab1p that is mediated through a translation factor eIF4G in the yeast *S. cerevisiae* (Wells et al., 1998). Using Atomic Force Microscopy, they further demonstrated that RNAs bound to this complex circularize and the circularization is lost upon disruption of this complex. To test if these interactions exist *in vivo*, Archer *et al*. performed immunoprecipitations with either a cap-binding protein (eIF4E), poly-A binding protein (PABP) or the adaptor protein (eIF4G) and probed for the presence of the other proteins after RNase treatment (Archer et al., 2015). The rationale being, if the interactions exist *in vivo*, the proteins would co-purify in spite of the RNase treatment digesting away the middle of the target mRNA. They designed qPCR primers to different regions of the mRNA to quantitatively measure these interactions. The data demonstrated that, in yeast, these c interactions do exist *in vivo*, however, they are present in a distinct phase of polysome assembly. Further, the level of these interactions varied based on the mRNA being probed.

This raised the question whether mammalian mRNAs adopt this configuration immediately after synthesis and splicing and are they transported to the cytoplasm with this arrangement? Or are RNAs packaged in a different conformation and adopt the circular configuration only upon engaging the translational machinery.

In our EJC RIPPLiT data, we detected very few interactions between the 5’ and 3' ends for more than 450 mRNA (transcripts with at least 100 chimeric junctions in at least one replicate). This suggested that the 5' and 3' of spliced, pre-translational mRNPs do not interact. However, it is possible that these interactions happen during translation. Thus, it would be interesting to apply RIPPLiT to ribosome associated mRNAs to test if mRNAs adopt a circular conformation during translation.

Capturing the higher order structure of mRNPs throughout their life cycle

In the current approach, we captured the higher order structures of mRNPs that have been spliced but released from chromatin and which have not yet been translated. We enrich the post-splicing and chromatin released complexes by spinning down our whole cell extracts at 15000 x g before performing the 1st IP. While the pre-translational mRNPs are captured by tandem IPs with the EJC, since EJCs are removed during the first round of translation by ribosomes. With this initial approach we were able to prove that EJC RIPPLiT can indeed capture mRNP higher order structure. However, mRNAs shed and gain a lot of different proteins during their life cycle (Moore, 2005)- from when they are in the nucleoplasm, till they are degraded by the exosome. A recent preprint in fact demonstrated that even the composition of the exon junction complex upon during the life cycle of an mRNP (Mabin et al., 2018). The SR-protein rich EJC associated

mRNAs also associate with the RNSP1 a peripheral protein of EJC complex. However, in the cytoplasm before translation the mRNP undergoes a change in composition with loss of SR-proteins and RNSP1 and gain of CASC3.

Thus, the next step in elucidating the dynamicity in mRNP structure would be to isolate mRNPs from different cellular fractions- chromatin associated, nucleoplasmic and cytoplasmic and measuring the changes, if any, in the chimeric junction frequency. One major challenge in doing this would the yield of mRNPs obtained through each fraction would undoubtedly be limited.

Capturing mRNP higher order organization within membrane-less organelles

Inside cells RNAs and their associated proteins form multiple different supramolecular structures both in the nucleus as well as the cytoplasm to assist their functions. These RNP bodies include the membrane-less organelles like nucleoli, paraspeckles, Cajal bodies in the nucleus, and processing bodies (P-bodies) and stress granules (SGs) in the cytoplasm (Banani et al., 2017). These aggregates of RNPs perform important functions in the life cycle of the RNAs like rRNP biogenesis (nucleoli), snRNP, snoRNP and other small RNP biogenesis (Cajal bodies), translationally repressed mRNA storages (P-bodies and SGs) (Decker and Parker, 2012; Stanek and Fox, 2017). These bodies can regulate the availability of the RNAs, control the rate of exchange of the RNAs and act as sponges for the RBPs. However, since these organelles are very dynamic, it has

proved difficult to identify the components of these organelles. Protocols have been developed to isolate these membrane-less organelles (e.g., (Hubstenberger et al., 2017) for purification of P-bodies). RIPPLiT could be applied to these purified organelles to understand the organization of the component RNPs. For instance, P-bodies were shown to contain hundreds of proteins and thousands of RNAs (Hubstenberger et al., 2017) demonstrating the complexity of these organelles. Since RIPPLiT involves ligating pieces of spatially proximal RNA fragments, they would capture the intra- as well as inter-RNA interactions present in these granules. Thus, this would aid in elucidating the 3D organization of these complex cellular bodies.

Capturing mRNP higher order organization inside specialized cells

An important mode of regulating cellular functions is by asymmetrical mRNA distribution and spatially compartmentalizing protein synthesis. One of the best studied examples are in neurons where the nuclei of these cells can be long distances away from the axons and dendrites (Glock et al., 2017). Further, they localize specific mRNAs to dendrites and axons to respond instantaneously to certain stimuli by making specific proteins upon receiving the signal. By having specific mRNAs ready for translation at a given dendrite or axon gives the cell the ability to process received signal in a temporally and spatially compartmentalized manner. More than 2,500 mRNAs have been shown to be locally translated in

different types of neurons (Glock et al., 2017). These localized mRNAs have been shown to have alternative 3' UTRs and poly adenylation sites. Thus, these could provide binding sites for different RNA-binding proteins affecting their localization. Further it would be interesting to study any 3D organizational differences in these transcripts versus the ones that are not localized. Protocols have been established to isolate neuronal processes and cell bodies to study mRNA distribution. These could be instead combined with RIPPLiT protocol to understand the 3D organization of these mRNPs. Furthermore, these mRNAs need to be packaged in a way that they are maintained in a silent state during transport and delivery and only translated upon receiving the cognate stimuli. Thus, it would be informative to study these mRNPs outside their localization sites and at the proper destination with and without stimuli.

What is the protein backbone around which mRNAs are organized?

The homogeneous interaction maps (**Figures 2.9 and 2.11A**) provide compelling evidence for the absence of a reproducible folding pattern with respect to mRNA sequence. The interpretation is that the mRNP is composed of many different proteins, only some of which are EJCs.

To test this further, we made a pile up plot heatmap by overlaying all heatmaps and plotting the interaction frequency 25 bins (20 nt X 20 nt) around the splice junctions (**Figure 4.2**). The heatmap was normalized by iterative correction

(Imakaev et al., 2012). If there was any bias due to EJC binding (~24 nt upstream of an exon-exon junction), the interaction frequency would be lower in the top right and bottom left corners. However, we see no such decrease in signal.

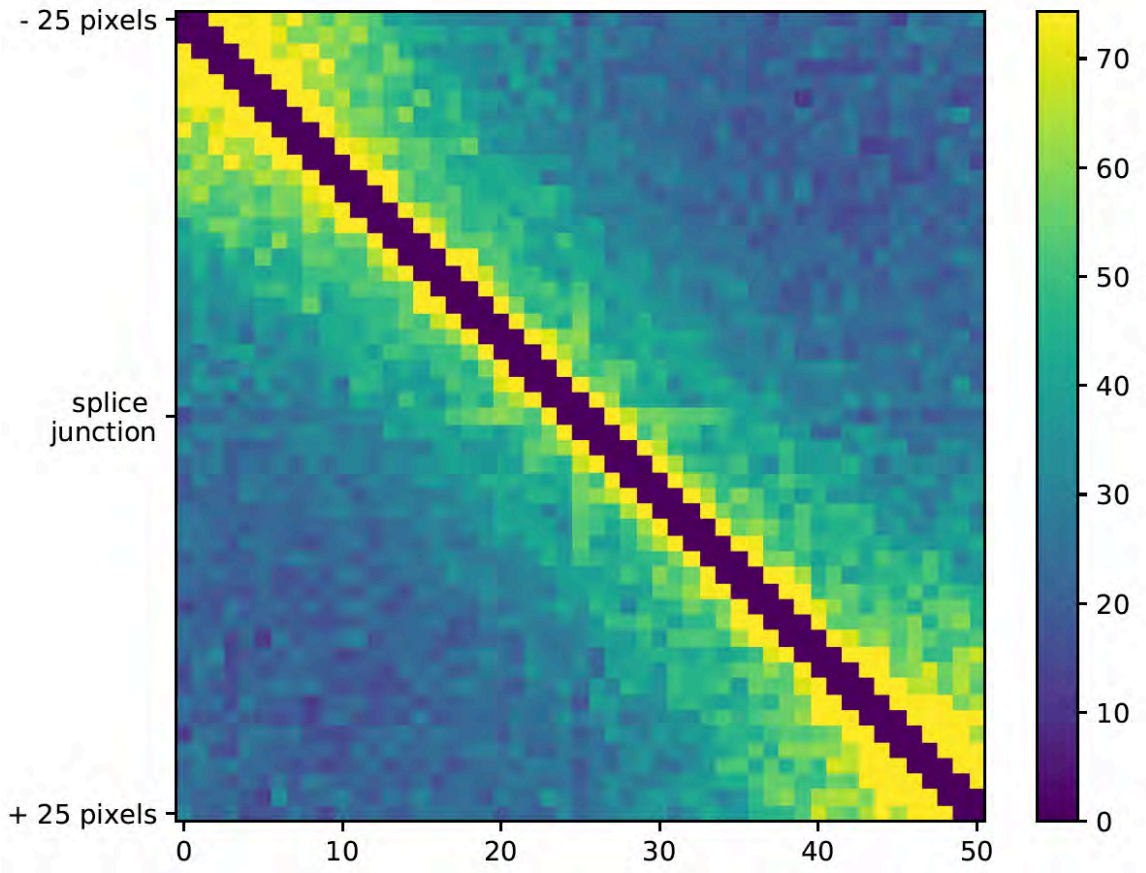


Figure 4.3. Chimeric junctions are not biased by the EJC

Heatmaps for the pileup around splice junctions (25 pixel on each side of the annotated splice junction). Heatmap was binned at 20 X 20 nts. Color scale indicates chimeric junction frequency and the heatmaps were normalized using iterative correction.

The existence of footprints in an ensemble population does not necessitate that every individual molecule within that population have exactly the same footprint pattern or occupancy. Indeed, in our lab's 2012 paper (Singh et al., 2012), we showed that even on the most abundant mRNAs only ~80% of potential cEJC deposition sites (i.e., 24 nts upstream of exon-exon junctions) are occupied (Figure 4 in that paper). Further, the relative heights of cEJC peaks upstream of different exon-exon junctions within the same mRNA species were highly variable (Figure 4A in that paper). Taken together, these findings suggest that the exact pattern of EJC deposition on any one mRNA molecule is a consequence of many stochastic events (e.g., whether or not the region 24 nts upstream of a 5' splice site was in a single stranded state at the moment when eIF4AIII and the Magoh/Y14 heterodimer come into close proximity on the spliceosome; whether or not an SR protein is bound nearby to provide additional stabilizing interactions at the time of EJC deposition). Therefore, the exact pattern of protein-RNA and protein-protein interactions will differ between individual mRNP molecules each containing the same mRNA species.

We did perform multiple analyses to assess the relationship between EJC peak height in our 2012 study (RIPiT) and ligation junction frequency in the current study (RIPPLiT). In no case, however, did we observe any statistically significant relationship. A major difference between the two studies were the RNA fragment sizes. Whereas the RIPiT libraries contained ~30 nt fragments produced upon highly stringent RNase 1 digestion, fragments in the RIPPLiT libraries range from

200 to 400 nts produced by mild RNase T1 digestion. As a result, there is almost no relationship between RIPiT and RIPPLiT fragment coverage at nucleotide resolution (compare the beige and red tracks in **Figure 2.10**). Thus, the absence of correlation between RIPiT EJC peak height and RIPPLiT ligation junction frequency is to be expected. This supports the model that folding of each individual mRNA molecule is different.

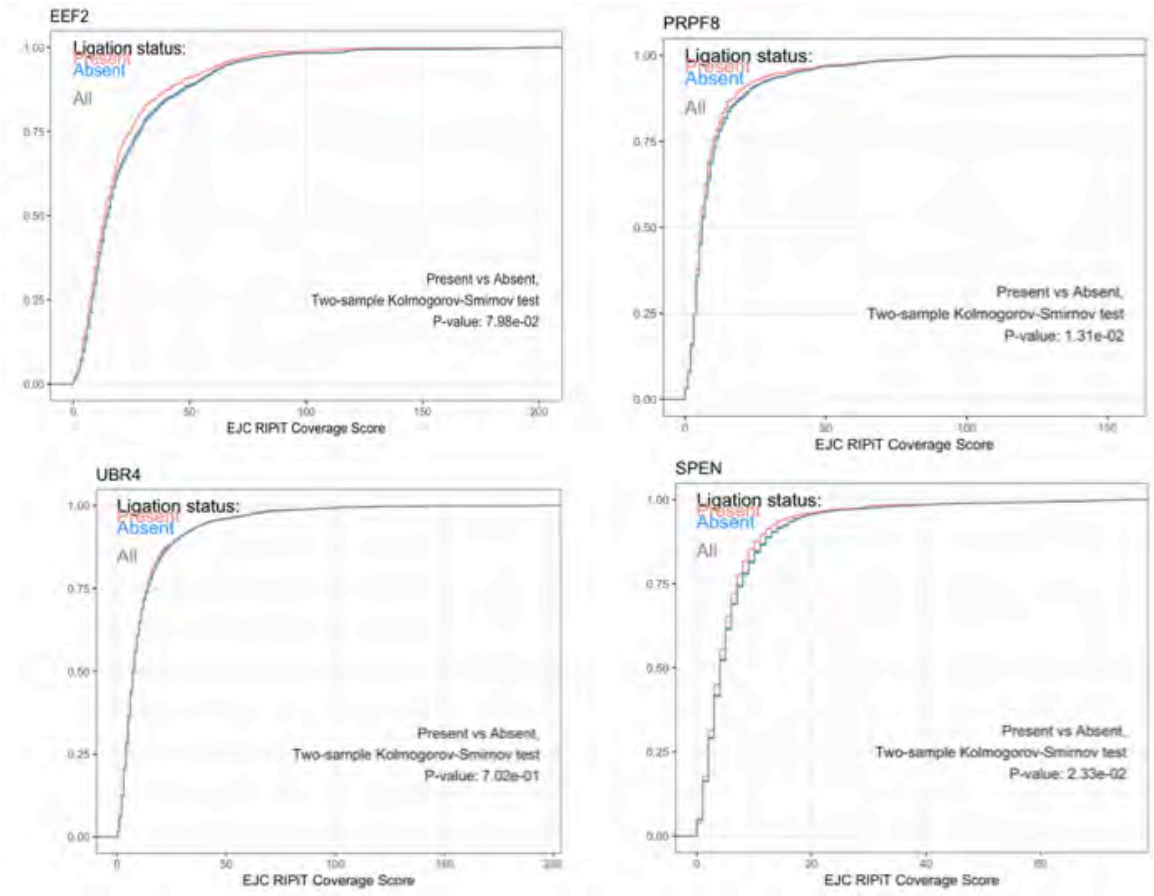


Figure 4.4. No significant difference in RIPiT coverage score for nucleotides present or absent in chimeric junctions

Cumulative frequency distribution of RIPiT coverage scores for all nucleotides (gray line), nucleotides present (salmon line) or absent (blue line) in chimeric junctions. Data for 4 genes is shown.

This also demonstrates that EJCs were simply used to pull out the mRNPs, but the structure is mediated by many different proteins. Proteins like YB-1 were shown to coat mRNAs throughout their length (Skabkin et al., 2004) and so could potentially form the backbone. Other abundant proteins associated with mRNAs are SR-proteins and hnRNP proteins (Busch and Hertel, 2012). Previous data from our lab has shown that SR-proteins indeed interact with EJCs to form megadalton sized complexes (Singh et al., 2012). To comprehensively identify proteins that bind an mRNA, Luhrmann lab affinity purified and spliced specific mRNAs (Merz et al., 2007). They then performed mass spectroscopy to ascertain the set of proteins pulled down with these mRNAs. Using this approach, they were able to identify ~45 proteins that bind a spliced mRNA. This list of proteins would provide a good starting point in elucidating the backbone of mRNA organization. Further, it would be interesting to apply RIPPLiT to these proteins, along with their systematic knockdown and over-expression, to obtain chimeric junction heatmap maps and facilitate identification of the protein backbone.

Improvements to ChimeraTie

A major effort of this project involved the development of a new analysis pipeline, ChimeraTie. Overall, using Bowtie2 (Langmead and Salzberg, 2012) is called iteratively in local alignment mode to map all fragments within a read. These are then put-together to identify pairwise interactions between each abutting fragment of a chimeric read. Specifically for EJC RIPPLiT, since our main focus

was mRNAs, the entire pipeline was optimized for the analysis of the transcriptome. However, ChimeraTie is universal in the sense that it can identify any type of chimeric junction like the ones created by circular RNAs, alternative splicing or even fusion transcripts. For identifying many of these events, one would need to use it to map it to the genome. While ChimeraTie was designed with the capabilities to do so, it has not been rigorously tested for genome mapping purposes. Thus, one of the next steps would be to test ChimeraTie with respect to genome mapping.

Analysis of higher order chimeric reads

Currently, ChimeraTie analyzes and visualizes pairwise interactions within reads. However, since EJC RIPPLiT is an ensemble assay and we are analyzing only pairwise interactions, it is difficult to predict 3D shapes of a single molecule of an mRNA. Interestingly, EJC RIPPLiT has 10,000s of reads **(Table 2.2**) with more than 2 fragments ligated together within the same read. These are important because these most likely signify interactions between regions of a single molecule. Thus, using these reads, we can start to construct a 3D model of individual molecules. However, ChimeraTie does not have the capabilities to handle these higher order interactions as well as any way to visualized them. This would be an interesting long-term avenue to pursue in the development of ChimeraTie pipeline.

Conclusion

In conclusion, with this thesis I have attempted to establish a complete toolkit, biochemical approach and a bioinformatics pipeline, for the capture and analysis of intra- and inter-RNA interactions within mRNP complexes. Applying this toolkit on EJC associated RNAs allowed us to, for the first time, envision the higher order structure of pre-translational mRNAs. Further, by performing polymer analysis on hundreds of mRNAs, we were able to propose the *in vivo* architecture of mRNAs packaged within mRNPs. Our data suggests that mRNAs are linearly and densely compacted into flexible rod-like structures before they undergo translation.

I envision that RIPPLiT could potentially be applied to any RNP of interest and thus can be universally applicable. Similarly, ChimeraTie can be used to analyze chimeric reads obtained from any experimental set-up. Thus, I predict it could help discover new biological phenomena in these set-ups that were previously overlooked due to unavailability of a proper tool. To this end, the guidelines provided in this thesis may motivate the community to use these tools and help improve them in the future.

REFERENCES

Adivarahan, S., Rahman, S., and Zenklusen, D. (2017). Spatial organization of single mRNPs at different stages of the gene expression pathway. bioRxiv.

Andersen, C.B., Ballut, L., Johansen, J.S., Chamieh, H., Nielsen, K.H., Oliveira, C.L., Pedersen, J.S., Seraphin, B., Le Hir, H., and Andersen, G.R. (2006). Structure of the exon junction core complex with a trapped DEAD-box ATPase bound to RNA. Science *313*, 1968-1972.

Archer, S.K., Shirokikh, N.E., Hallwirth, C.V., Beilharz, T.H., and Preiss, T. (2015). Probing the closed-loop model of mRNA translation in living cells. RNA Biol *12*, 248-254.

Ash, R.B. (1990). Information theory (Courier Corporation, 1990).

Aw, J.G., Shen, Y., Wilm, A., Sun, M., Lim, X.N., Boon, K.L., Tapsin, S., Chan, Y.S., Tan, C.P., Sim, A.Y.*, et al.* (2016). In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. Mol Cell *62*, 603-617.

Bajad, P., Jantsch, M.F., Keegan, L., and O'Connell, M. (2017). A to I editing in disease is not fake news. RNA Biol *14*, 1223-1231.

Banani, S.F., Lee, H.O., Hyman, A.A., and Rosen, M.K. (2017). Biomolecular condensates: organizers of cellular biochemistry. Nat Rev Mol Cell Biol *18*, 285-298.

Batisse, J., Batisse, C., Budd, A., Bottcher, B., and Hurt, E. (2009). Purification of nuclear poly(A)-binding protein Nab2 reveals association with the yeast transcriptome and a messenger ribonucleoprotein core structure. J Biol Chem *284*, 34911-34917.

Bono, F., Cook, A.G., Grunwald, M., Ebert, J., and Conti, E. (2010). Nuclear import mechanism of the EJC component Mago-Y14 revealed by structural studies of importin 13. Mol Cell *37*, 211-222.

Brennan, C.M., and Steitz, J.A. (2001). HuR and mRNA stability. Cell Mol Life Sci *58*, 266-277.

Busch, A., and Hertel, K.J. (2012). Evolution of SR protein and hnRNP splicing regulatory factors. Wiley Interdiscip Rev RNA *3*, 1-12.

Carson, J.H., Gao, Y., Tatavarty, V., Levin, M.K., Korza, G., Francone, V.P., Kosturko, L.D., Maggipinto, M.J., and Barbarese, E. (2008). Multiplexed RNA trafficking in oligodendrocytes and neurons. Biochim Biophys Acta *1779*, 453-458.

Celotto, A.M., and Graveley, B.R. (2001). Alternative Splicing of the Drosophila Dscam Pre-mRNA Is Both Temporally and Spatially Regulated. GENETICS *vol. 159* 599-608.

Christensen, A.K., Kahn, L.E., and Bourne, C.M. (1987). Circular polysomes predominate on the rough endoplasmic reticulum of somatotropes and mammotropes in the rat anterior pituitary. Am J Anat *178*, 1-10.

Conway, G., Wooley, J., Bibring, T., and LeStourgeon, W.M. (1988). Ribonucleoproteins package 700 nucleotides of pre-mRNA into a repeating array of regular particles. Molecular and Cellular Biology *8*, 2884-2895.

Crick, F. (1958). On Protein Synthesis. The Symposia of the Society for Experimental Biology *12*, 138-163.

de Wit, E., and de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. Genes Dev *26*, 11-24.

Decker, C.J., and Parker, R. (2012). P-bodies and stress granules: possible roles in the control of translation and mRNA degradation. Cold Spring Harb Perspect Biol *4*, a012286.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306-1311.

Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C., and Assmann, S.M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature *505*, 696-700.

Dobin, A., and Gingeras, T.R. (2015). Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics *51*, 11 14 11-19.

Dostie, J., and Dreyfuss, G. (2002). Translation is required to remove Y14 from mRNAs in the cytoplasm. Current Biology *12*, 1060-1067.

Dreyfuss, G. (1986). Structure and function of nuclear and cytoplasmic ribonucleoprotein particles. Annu Rev Cell Biol *2*, 459-498.

Ferrandon, D., Elphick, L., Nüsslein-Volhard, C., and St Johnston, D. (1994). Staufen protein associates with the 3′UTR of bicoid mRNA to form particles that move in a microtubule-dependent manner. Cell *79*, 1221-1232.

Ferrandon, D., Koch, I., Westhof, E., and Nusslein-Volhard, C. (1997). RNA-RNA interaction is required for the formation of specific bicoid mRNA 3' UTR-STAUFEN ribonucleoprotein particles. EMBO J *16*, 1751-1758.

Fresno, M., Jimenez, A., and Vazquez, D. (1977). Inhibition of Translation in Eukaryotic Systems by Harringtonine. European Journal of Biochemistry *72*, 323-330.

Fudenberg, G., and Mirny, L.A. (2012). Higher-order chromatin structure: bridging physics and biology. Curr Opin Genet Dev *22*, 115-124.

Gatenby, R.A., and Frieden, B.R. (2007). Information theory in living systems, methods, applications, and challenges. Bull Math Biol *69*, 635-657.

Ge, Z., Quek, B.L., Beemon, K.L., and Hogg, J.R. (2016). Polypyrimidine tract binding protein 1 protects mRNAs from recognition by the nonsense-mediated mRNA decay pathway. Elife *5*.

Gehring, N.H., Lamprinaki, S., Kulozik, A.E., and Hentze, M.W. (2009). Disassembly of exon junction complexes by PYM. Cell *137*, 536-548.

Gibcus, J.H., Samejima, K., Goloborodko, A., Samejima, I., Naumova, N., Nuebler, J., Kanemaki, M.T., Xie, L., Paulson, J.R., Earnshaw, W.C.*, et al.* (2018). A pathway for mitotic chromosome formation. Science *359*.

Glock, C., Heumuller, M., and Schuman, E.M. (2017). mRNA transport & local translation in neurons. Curr Opin Neurobiol *45*, 169-177.

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell *153*, 654-665.

Hinde, E., Thammasiraphop, K., Duong, H.T., Yeow, J., Karagoz, B., Boyer, C., Gooding, J.J., and Gaus, K. (2017). Pair correlation microscopy reveals the role of nanoparticle shape in intracellular transport and site of drug release. Nat Nanotechnol *12*, 81-89.

Hoffmann, S., Otto, C., Doose, G., Tanzer, A., Langenberger, D., Christ, S., Kunz, M., Holdt, L.M., Teupser, D., Hackermuller, J.*, et al.* (2014). A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. Genome Biol *15*, R34.

Hubstenberger, A., Courel, M., Benard, M., Souquere, S., Ernoult-Lange, M., Chouaib, R., Yi, Z., Morlot, J.B., Munier, A., Fradet, M.*, et al.* (2017). P-Body Purification Reveals the Condensation of Repressed mRNA Regulons. Mol Cell *68*, 144-157 e145.

Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods *9*, 999-1003.

Katahira, J. (2015). Nuclear export of messenger RNA. Genes (Basel) *6*, 163-184.

Khatter, H., Myasnikov, A.G., Natchiar, S.K., and Klaholz, B.P. (2015). Structure of the human 80S ribosome. Nature *520*, 640-645.

Khong, A., and Parker, R. (2018). mRNP architecture in translating and stress conditions reveals an ordered pathway of mRNP compaction. BioRxiv.

Kinniburgh, A.J., and Martin, T.E. (1976). Detection of mRNA sequences in nuclear 30S ribonucleoprotein subcomplexes. Proceedings of the National Academy of Sciences *73*, 2725-2729.

Kotlarz, D., Fritsch, A., and Buc, H. (1986). Variations of intramolecular ligation rates allow the detection of protein-induced bends in DNA. EMBO J *5*, 799-803.

Kramer, M.C., and Gregory, B.D. (2018). Does RNA secondary structure drive translation or vice versa? Nat Struct Mol Biol *25*, 641-643.

Kudla, G., Granneman, S., Hahn, D., Beggs, J.D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. Proc Natl Acad Sci USA *108*, 10010-10015.

Lai, D., Proctor, J.R., and Meyer, I.M. (2013). On the importance of cotranscriptional RNA structure formation. Rna *19*, 1461-1473.

Lajoie, B.R., Dekker, J., and Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods *72*, 65-75.

Landau, H.B., Maddock, J.T., Shoemaker, F.F., and Costello, J.G. (1982). An information transfer model to define information users and outputs with specific application to environmental technology. Journal of the American Society for Information Science *33*, 82-91.

Langdon, E.M., Qiu, Y., Ghanbari Niaki, A., McLaughlin, G.A., Weidmann, C.A., Gerbich, T.M., Smith, J.A., Crutchley, J.M., Termini, C.M., Weeks, K.M.*, et al.* (2018). mRNA structure determines specificity of a polyQ-driven phase separation. Science *360*, 922-927.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Le Hir, H., Moore, M.J., and Maquat, L.E. (2000). Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon–exon junctions. Genes & Dev *14*, 1098-1108.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843-854.

Lejeune, F., Ishigaki , Y., Li, X., and Maquat, L.E. (2002). The exon junction complex is detected on CBP80-bound but not eIF4E-bound mRNA in mammalian cells: dynamics of mRNP remodeling. The EMBO Journal *21*, 3536-3545.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, F., Zheng, Q., Vandivier, L.E., Willmann, M.R., Chen, Y., and Gregory, B.D. (2012). Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. Plant Cell *24*, 4346-4359.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, S., Xu, Z., and Sheng, J. (2018). tRNA-Derived Small RNA: A Novel Regulatory Small Non-Coding RNA. Genes (Basel) *9*.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O.*, et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289-293.

Liu, F., and Gong, C.X. (2008). Tau exon 10 alternative splicing and tauopathies. Mol Neurodegener *3*, 8.

Lu, Z., Zhang, Q.C., Lee, B., Flynn, R.A., Smith, M.A., Robinson, J.T., Davidovich, C., Gooding, A.R., Goodrich, K.J., Mattick, J.S.*, et al.* (2016). RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. Cell *165*, 1267-1279.

Mabin, J.W., Woodward, L.A., Patton, R., Yi, Z., Jia, M., Wysocki, V., Bundschuh, R., and Singh, G. (2018). The exon junction complex undergoes a compositional switch that alters mRNP structure and nonsense-mediated mRNA decay activity. bioRxiv.

Malcolm, D.B., and Sommerville, J. (1977). The structure of nuclear ribonucleoprotein of amphibian oocytes. Journal of Cell Science *24*, 143-165.

Marini, J.C., Levene, S.D., Crothers, D.M., and Englund, P.T. (1982). Bent helical structure in kinetoplast DNA. Proceedings of the National Academy of Sciences *79*, 7664-7668.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal *17*, 10.

Matsumoto, K., Tanaka, K.J., Aoki, K., Sameshima, M., and Tsujimoto, M. (2003). Visualization of the reconstituted FRGY2–mRNA complexes by electron microscopy. Biochemical and Biophysical Research Communications *306*, 53-58.

McManus, C.J., and Graveley, B.R. (2011). RNA structure and the mechanisms of alternative splicing. Curr Opin Genet Dev *21*, 373-379.

Merz, C., Urlaub, H., Will, C.L., and Luhrmann, R. (2007). Protein composition of human mRNPs spliced in vitro and differential requirements for mRNP protein recruitment. RNA *13*, 116-128.

Mishler, D.M., Christ, A.B., and Steitz, J.A. (2008). Flexibility in the site of exon junction complex deposition revealed by functional group and RNA secondary structure alterations in the splicing substrate. RNA *14*, 2657-2670.

Miura, F., Kawaguchi, N., Yoshida, M., Uematsu, C., Kito, K., Sakaki, Y., and Ito, T. (2008). Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. BMC Genomics *9*, 574.

Moore, M.J. (2005). From birth to death: the complex lives of eukaryotic mRNAs. Science *309*, 1514-1518.

Mor, A., Suliman, S., Ben-Yishay, R., Yunger, S., Brody, Y., and Shav-Tal, Y. (2010). Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells. Nat Cell Biol *12*, 543-552.

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie , B.R., Mirny, L.A., and Dekker, J. (2013). Organization of the Mitotic Chromosome. Science *Vol. 342*, 948-953.

Ng, K., Pullirsch, D., Leeb, M., and Wutz, A. (2007). Xist and the order of silencing. EMBO Rep *8*, 34-39.

Nguyen, T.C., Cao, X., Yu, P., Xiao, S., Lu, J., Biase, F.H., Sridhar, B., Huang, N., Zhang, K., and Zhong, S. (2016). Mapping RNA-RNA interactome and RNA structure in vivo by MARIO. Nat Commun *7*, 12023.

Palazzo, A.F., and Lee, E.S. (2015). Non-coding RNA: what is functional and what is junk? Front Genet *6*, 2.

Park, H.Y., Lim, H., Yoon, Y.J., Follenzi, A., Nwokafor, C., Lopez-Jones, M., Meng, X., and Singer, R.H. (2014). Visualization of dynamics of single endogenous mRNA labeled in live mouse. Science *343*, 422-424.

Piao, M., Sun, L., and Zhang, Q.C. (2017). RNA Regulations and Functions Decoded by Transcriptome-wide RNA Structure Probing. Genomics Proteomics Bioinformatics *15*, 267-278.

Piccinelli, P., and Samuelsson, T. (2007). Evolution of the iron-responsive element. RNA *13*, 952-966.

Pintacuda, G., Young, A.N., and Cerase, A. (2017). Function by Structure: Spotlights on Xist Long Non-coding RNA. Front Mol Biosci *4*, 90.

Ramani, V., Qiu, R., and Shendure, J. (2015). High-throughput determination of RNA structure by proximity ligation. Nat Biotechnol *33*, 980-984.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S.*, et al.* (2014).

A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665-1680.

Ricci, E.P., Kucukural, A., Cenik, C., Mercier, B.C., Singh, G., Heyer, E.E., Ashar-Patel, A., Peng, L., and Moore, M.J. (2014). Staufen1 senses overall transcript secondary structure to regulate translation. Nat Struct Mol Biol *21*, 26-35.

Rivas, E., Klein, R.J., Jones, T.A., and Eddy, S.R. (2001). Computational identification of noncoding RNAs in E. coli by comparative genomics. Current Biology *11*, 1369-1373.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J.S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature *505*, 701-705.

Samarina, O.P., Lukanidin, E.M., Molnar, J., and Georgiev, G.P. (1968). Structural organization of nuclear complexes containing DNA-like RNA. Journal of Molecular Biology *33*, 251-263.

Sauliere, J., Murigneux, V., Wang, Z., Marquenet, E., Barbosa, I., Le Tonqueze, O., Audic, Y., Paillard, L., Roest Crollius, H., and Le Hir, H. (2012). CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. Nat Struct Mol Biol *19*, 1124-1131.

Sharma, E., Sterne-Weiler, T., O'Hanlon, D., and Blencowe, B.J. (2016). Global Mapping of Human RNA-RNA Interactions. Mol Cell *62*, 618-626.

Shibuya, T., Tange, T.O., Sonenberg, N., and Moore, M.J. (2004). eIF4AIII binds spliced mRNA in the exon junction complex and is essential for nonsense-mediated decay. Nat Struct Mol Biol *11*, 346-351.

Singh, G., Kucukural, A., Cenik, C., Leszyk, J.D., Shaffer, S.A., Weng, Z., and Moore, M.J. (2012). The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. Cell *151*, 750-764.

Singh, G., Pratt, G., Yeo, G.W., and Moore, M.J. (2015). The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion. Annu Rev Biochem *84*, 325-354.

Singh, G., Ricci, E.P., and Moore, M.J. (2014). RIPiT-Seq: a high-throughput approach for footprinting RNA:protein complexes. Methods *65*, 320-332.

Skabkin, M.A., Kiselyova, O.I., Chernov, K.G., Sorokin, A.V., Dubrovin, E.V., Yaminsky, I.V., Vasiliev, V.D., and Ovchinnikov, L.P. (2004). Structural organization of mRNA complexes with major core mRNP protein YB-1. Nucleic Acids Res *32*, 5621-5635.

Skoglund, U., Andersson, K., Björkroth, B., Lamb, M., and Daneholt, B. (1983). Visualization of the formation and transport of a specific hnRNP particle. Cell *34*, 847-855.

Smola, M.J., Christy, T.W., Inoue, K., Nicholson, C.O., Friedersdorf, M., Keene, J.D., Lee, D.M., Calabrese, J.M., and Weeks, K.M. (2016). SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. Proc Natl Acad Sci USA *113*, 10322-10327.

Solomon, O., Di Segni, A., Cesarkas, K., Porath, H.T., Marcu-Malina, V., Mizrahi, O., Stern-Ginossar, N., Kol, N., Farage-Barhom, S., Glick-Saar, E.*, et al.* (2017). RNA editing by ADAR1 leads to context-dependent transcriptome-wide changes in RNA secondary structure. Nat Commun *8*, 1440.

Spitale, R.C., Flynn, R.A., Zhang, Q.C., Crisalli, P., Lee, B., Jung, J.W., Kuchelmeister, H.Y., Batista, P.J., Torre, E.A., Kool, E.T.*, et al.* (2015). Structural imprints in vivo decode RNA regulatory mechanisms. Nature *519*, 486-490.

Stanek, D., and Fox, A.H. (2017). Nuclear bodies: news insights into structure and function. Curr Opin Cell Biol *46*, 94-101.

Sugimoto, Y., Vigilante, A., Darbo, E., Zirra, A., Militti, C., D'Ambrogio, A., Luscombe, N.M., and Ule, J. (2015). hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. Nature *519*, 491-494.

Teng, L., He, B., Wang, J., and Tan, K. (2015). 4DGenome: a comprehensive database of chromatin interactions. Bioinformatics *31*, 2560-2564.

Tian, N., Yang, Y., Sachsenmaier, N., Muggenhumer, D., Bi, J., Waldsich, C., Jantsch, M.F., and Jin, Y. (2011). A structural determinant required for RNA editing. Nucleic Acids Res *39*, 5669-5681.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Travis, A.J., Moody, J., Helwak, A., Tollervey, D., and Kudla, G. (2014). Hyb: a bioinformatics pipeline for the analysis of CLASH (crosslinking, ligation and sequencing of hybrids) data. Methods *65*, 263-273.

Van Treeck, B., Protter, D.S.W., Matheny, T., Khong, A., Link, C.D., and Parker, R. (2018). RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome. Proc Natl Acad Sci USA *115*, 2734-2739.

Viollet, S., Fuchs, R.T., Munafo, D.B., Zhuang, F., and Robb, G.B. (2011). T4 RNA ligase 2 truncated active site mutants: improved tools for RNA analysis. BMC Biotechnol *11*, 72.

Wang, J., Yang, Y., Yu, M., Hu, G., Gan, Y., Gao, H., and Shi, X. (2018). Diffusion of rod-like nanoparticles in non-adhesive and adhesive porous polymeric gels. Journal of the Mechanics and Physics of Solids *112*, 431-457.

Watters, K.E., Strobel, E.J., Yu, A.M., Lis, J.T., and Lucks, J.B. (2016). Cotranscriptional folding of a riboswitch at nucleotide resolution. Nat Struct Mol Biol *23*, 1124-1131.

Wells, S.E., Hillner, P.E., Vale, R.D., and Sachs, A.B. (1998). Circularization of mRNA by Eukaryotic Translation Initiation Factors. Molecular Cell *2*, 135-140.

Woodward, L.A., Mabin, J.W., Gangras, P., and Singh, G. (2017). The exon junction complex: a lifelong guardian of mRNA fate. Wiley Interdiscip Rev RNA *8*.

Wutz, A., Rasmussen, T.P., and Jaenisch, R. (2002). Chromosomal silencing and localization are mediated by different domains of Xist RNA. Nat Genet *30*, 167-174.

Yue, M., and Ogawa, Y. (2018). CRISPR/Cas9-mediated modulation of splicing efficiency reveals short splicing isoform of Xist RNA is sufficient to induce X-chromosome inactivation. Nucleic Acids Res *46*, e26.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics *30*, 614-620.

Zhuang, F., Fuchs, R.T., Sun, Z., Zheng, Y., and Robb, G.B. (2012). Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. Nucleic Acids Res *40*, e54.